

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

DMAP : UNE NOUVELLE MÉTHODE DE CARTOGRAPHIE GÉNÉTIQUE
FINE ADAPTÉE À DES MODÈLES GÉNÉTIQUES COMPLEXES

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES

PAR
MARIE-HÉLÈNE DESCARY

AOÛT 2012

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.01-2006). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

J'aimerais tout d'abord remercier Fabrice Larribe, mon directeur de recherche. Un immense merci pour ces deux belles années à travailler sous ta supervision. Merci pour ton soutien, ta patience, tes précieux conseils et pour la confiance que tu m'as accordée.

Je veux de plus remercier Sorana Froda pour sa générosité et son aide précieuse de tous les jours.

Je remercie également le corps professoral de l'UQAM. Merci pour votre dévouement et pour la passion que vous mettez dans votre travail.

Merci à ma famille et à mes amis qui m'ont toujours soutenue et encouragée pendant la réalisation de ce mémoire.

Finalement, un merci tout particulier à Guillaume. Merci pour ta compréhension, ton écoute et ta patience.

TABLE DES MATIÈRES

LISTE DES FIGURES	ix
LISTE DES TABLEAUX	xiii
RÉSUMÉ	xv
INTRODUCTION	1
CHAPITRE I	
INTRODUCTION À LA GÉNÉTIQUE	3
1.1 Définitions et explications des termes de base en génétique	3
1.1.1 Les chromosomes	3
1.1.2 L'ADN	4
1.1.3 Les gènes	4
1.2 Transmission du bagage génétique chez l'humain	6
1.3 Variabilité génétique	7
1.3.1 Sources de variation génétique	7
1.3.2 Distances et cartographie génétique	9
1.3.3 Marqueurs génétiques	10
CHAPITRE II	
THÉORIE DE LA COALESCENCE	13
2.1 Le modèle de Wright-Fisher	13
2.2 Le processus de coalescence	14
2.2.1 Coalescence à temps discret	16
2.2.2 Coalescence à temps continu	17
2.3 Processus de coalescence avec mutation	18
2.4 Le graphe de recombinaison ancestral	20
CHAPITRE III	
MÉTHODES DE CARTOGRAPHIE GÉNÉTIQUE FINE	27
3.1 Éléments communs aux trois méthodes	27
3.2 Méthode MapARG	28

3.2.1	Le modèle	28
3.2.2	La probabilité d'un graphe de recombinaison ancestral	30
3.2.3	Échantillonnage pondéré : deux distributions proposées d'intérêt	36
3.2.4	Vraisemblance composite et conditionnelle	41
3.2.5	Inférence d'allèle au marqueur correspondant au TIM	43
3.2.6	Exemple de résultat	45
3.3	Méthode Margarita	45
3.3.1	Le modèle	45
3.3.2	Graphe de recombinaison ancestral et arbre partiel	46
3.3.3	Inférence du TIM et algorithme utilisé par Margarita	50
3.3.4	Exemple de résultat	54
3.4	Méthode LATAG	54
3.4.1	Le modèle	55
3.4.2	Calcul de la probabilité du phénotype	57
3.4.3	Exemple de résultat	57
CHAPITRE IV		
UNE NOUVELLE MÉTHODE DE CARTOGRAPHIE GÉNÉTIQUE FINE		59
4.1	Idée générale	60
4.2	Le modèle	61
4.2.1	Calcul de la fonction de vraisemblance	61
4.2.2	Calcul de la probabilité du phénotype	63
4.3	Algorithme et défi computationnel	69
4.3.1	Algorithme de DMap	69
4.3.2	Défis computationnels de DMap	71
4.4	Test d'hypothèses	76
CHAPITRE V		
RÉSULTATS		81
5.1	Simulation des données	82
5.2	Résultats obtenus avec la méthode DMap	83

5.2.1	Les paramètres de simulation	83
5.2.2	Comparaison des différentes versions de DMap	85
5.2.3	Tests d'hypothèses	93
5.3	Comparaison de DMap avec trois autres méthodes de cartographie génétique	93
5.3.1	Méthode basée sur l'étude d'association	94
5.3.2	Illustration des résultats obtenus avec chaque méthode	95
5.3.3	Distribution des biais obtenus avec chaque méthode	105
5.4	Discussion	112
	CONCLUSION	113
	APPENDICE A	
	APPROXIMATION DE LA DISTRIBUTION CONDITIONNELLE $\pi(\cdot H)$	115
A.1	Interprétation généalogique	115
A.2	Calcul formel de $\hat{\pi}(\cdot H)$	116
A.3	Calcul récursif de $\hat{\pi}(h_{k+1} h_1, \dots, h_k)$	117
	APPENDICE B	
	ALGORITHME EN LANGAGE C++	119
	RÉFÉRENCES	125

LISTE DES FIGURES

Figure	Page
1.1 Deux représentations équivalentes d'une molécule d'ADN.	4
1.2 Diplotype versus génotype	5
1.3 Illustration d'une méiose	8
1.4 Illustration d'un double enjambement	8
1.5 Illustration des deux situations dans lesquelles le terme cartographie généétique est employé.	11
1.6 Conversion de données en SNPs	12
2.1 Exemple de l'évolution d'une population sous le modèle Wright-Fisher .	15
2.2 Représentation d'une généalogie	15
2.3 Illustration d'un événement de recombinaison	22
2.4 Illustration d'un graphe de recombinaison ancestral	24
2.5 Illustration d'un ARG généré à partir d'un échantillon donné.	26
3.1 Illustration d'un échantillon de séquences génétiques	29
3.2 Illustration des événements de transitions possibles dans un ARG	33
3.3 Représentation d'un échantillon de séquences génétiques	34
3.4 Illustration du passage de l'état H_τ à l'état $H_{\tau+1}$	37
3.5 Fenêtre de marqueurs	42
3.6 Exemple de résultat obtenu avec MapARG	46
3.7 Illustration de la simulation d'un événement de recombinaison avec la méthode Margarita	49
3.8 Illustration d'un arbre partiel	51
3.9 Illustration de l'ajout du TIM sur un arbre partiel	52

3.10 Exemple de résultat obtenu avec la méthode Margarita	54
3.11 Exemple de résultat obtenu avec LATAG	58
4.1 Résumé de la méthode DMap	68
4.2 Illustration de la notation utilisée afin de représenter d'une façon concise chaque événement possible d'un graphe	72
4.3 Illustration de quelques concepts de la théorie des graphes	73
4.4 Illustration de l'algorithme présenté à la section 4.3.2	77
4.5 Exemple de résultat obtenu avec la méthode DMap	80
5.1 Résultats obtenus pour le scénario A de la population 54.	88
5.2 Résultats obtenus pour le scénario B de la population 54.	89
5.3 Résultats obtenus pour le scénario C de la population 54.	90
5.4 Résultats obtenus pour le scénario D de la population 54.	91
5.5 Figure illustrant les distributions des biais des estimateurs obtenus avec les deux versions de la fonction de vraisemblance à comparer.	92
5.6 Résultats obtenus avec l'échantillon généré selon le scénario A de la po- pulation 58 pour les 4 différentes méthodes	97
5.7 Résultats obtenus avec l'échantillon généré selon le scénario B de la po- pulation 58 pour les 4 différentes méthodes	98
5.8 Résultats obtenus avec l'échantillon généré selon le scénario C de la po- pulation 58 pour les 4 différentes méthodes	99
5.9 Résultats obtenus avec l'échantillon généré selon le scénario D de la po- pulation 58 pour les 4 différentes méthodes	100
5.10 Résultats obtenus avec l'échantillon généré selon le scénario F de la po- pulation 105 pour les 4 différentes méthodes	101
5.11 Résultats obtenus avec l'échantillon généré selon le scénario G de la po- pulation 105 pour les 4 différentes méthodes	102
5.12 Résultats obtenus avec l'échantillon généré selon le scénario H de la po- pulation 105 pour les 4 différentes méthodes	103
5.13 Résultats obtenus avec l'échantillon généré selon le scénario I de la po- pulation 105 pour les 4 différentes méthodes	104

- 5.14 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario A. 108
- 5.15 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario B. 108
- 5.16 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario C. 109
- 5.17 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario D. 109
- 5.18 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario F. 110
- 5.19 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario G. 110
- 5.20 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario H. 111
- 5.21 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario I. 111

LISTE DES TABLEAUX

Tableau	Page
1.1 Lien entre le génotype et le phénotype du gène ABO.	6
3.1 Table de contingence obtenue suite à l'inférence illustrée à la figure 3.9 .	53
4.1 Tableau résumant les 4 catégories possibles pour une séquence	63
4.2 Exemple de résultats obtenus suite à l'ajout de mutations sur les branches d'un arbre partiel	69
5.1 Tableau résumant les caractéristiques des échantillons générés pour cha- cune des 200 populations simulées.	84
5.2 Tableau contenant les valeur-p obtenues suite au test d'hypothèse sur les 4 échantillons de la population 54.	93
5.3 Table de contingence utilisée pour des tests d'association entre une ma- ladie génétique et un marqueur donné.	95

RÉSUMÉ

Dans ce mémoire, nous présentons une nouvelle méthode de cartographie génétique fine ayant comme particularité de pouvoir être utilisée dans le cadre de modèles génétiques complexes. Nous présentons tout d'abord quelques concepts de génétique et de statistique génétique, avec une emphase particulière sur le processus de coalescence qui est à la base de notre travail. Par la suite, trois méthodes de cartographie génétique déjà existantes sont présentées; notre nouvelle méthode contient des éléments de chacune d'entre elles. Nous décrivons ensuite la nouvelle méthode proposée dans ce mémoire. Finalement, nous testons notre nouvelle approche à l'aide de simulations; nous comparons par la suite les résultats obtenus avec notre méthode à ceux obtenus par des tests d'association classiques, et par deux des trois méthodes présentées au début du mémoire. Les résultats nous laissent croire que notre méthode est performante autant dans des cas de modèles génétiques simples que complexes, contrairement à la plupart des méthodes existantes.

MOTS-CLÉS : cartographie génétique, processus de coalescence, arbre de recombinaison ancestral, arbre partiel, distribution proposée, fonction de pénétrance.

INTRODUCTION

Depuis quelques années, grâce à une technologie en constante évolution, le séquençage du génome humain est devenu une pratique couramment employée afin d'obtenir des données génétiques des plus complètes. L'accessibilité à ces immenses bases de données a favorisé le développement de plusieurs nouvelles méthodes statistiques afin d'extraire les informations pertinentes contenues dans ces données. Par exemple, dans les deux dernières décennies, plusieurs nouvelles méthodes de cartographie génétique ont vu le jour. Ces méthodes ont pour but d'identifier les gènes influençant des caractères d'intérêt sur un chromosome ; la connaissance de tels gènes est très importante à la compréhension des mécanismes des maladies génétiques et par le fait même, à la recherche de traitements contre celles-ci.

Le but de ce mémoire est de développer une nouvelle méthode de cartographie génétique fine, que nous avons nommé DMap, permettant d'estimer la position sur une séquence génétique d'une mutation influençant une maladie d'intérêt. La méthode DMap se distingue des méthodes déjà existantes par sa capacité à considérer des modèles génétiques complexes et par la rapidité de son exécution.

Le premier chapitre de ce mémoire contient une présentation des concepts de base en génétique qui seront essentiels à la compréhension des modèles statistiques présentés dans les chapitres subséquents. La théorie de la coalescence est présentée au second chapitre ; la méthode DMap s'appuie sur cette théorie afin de modéliser l'évolution des séquences génétiques dans le temps. Au chapitre trois, nous présentons trois méthodes de cartographie génétique fine qui nous ont inspirés lors de l'élaboration de la méthode DMap. Le chapitre quatre contient la description détaillée de la nouvelle méthode (DMap) que nous proposons dans ce mémoire. Finalement, le dernier chapitre contient la présentation de résultats obtenus avec des données simulées.

CHAPITRE I

INTRODUCTION À LA GÉNÉTIQUE

La génétique est une branche de la biologie qui étudie la transmission et la variation des caractères héréditaires chez les individus d'une même espèce. La connaissance des concepts de base de cette science, notamment des concepts de mutation et de recombinaison génétique, est essentielle à la compréhension des modèles statistiques présentés dans ce mémoire. Ce chapitre est donc consacré à l'introduction des termes de base en génétique et à l'étude de la transmission et des sources de diversité du bagage génétique chez l'humain. Le lecteur familier avec ce sujet est invité à passer directement au chapitre suivant.

1.1 Définitions et explications des termes de base en génétique

1.1.1 Les chromosomes

Un chromosome est une structure que l'on retrouve dans le noyau d'une cellule et qui est composé du matériel génétique des organismes : l'*acide désoxyribonucléique* (ADN). Notons que le nombre de chromosomes diffère d'un organisme à un autre ainsi que le nombre de copies de chaque chromosome : les organismes *haploïdes* (ayant un seul parent) ont une seule copie tandis que les *diploïdes* (ayant deux parents) en ont deux ; les chromosomes d'une même paire sont dits *homologues*. On appelle génome l'ensemble du matériel génétique contenu dans les chromosomes d'un organisme.

1.1.2 L'ADN

L'ADN est un acide nucléique composé de deux chaînes de nucléotides en forme d'hélice, où chaque nucléotide est caractérisé par une des quatre bases azotées suivantes : la cytosine (C), la thymine (T), l'adénine (A) et la guanine (G) (voir figure 1.1). Les nucléotides de chaque chaîne sont liés les uns aux autres par un pont d'hydrogène en respectant la règle suivante : la cytosine se lie à la guanine et l'adénine se lie à la thymine. Il est donc courant de représenter l'ADN par une seule chaîne de bases azotées, puisqu'on peut en déduire directement la composition de la seconde chaîne. L'ADN contient les instructions génétiques pour la synthèse des protéines, elle joue donc un rôle crucial dans le développement et le fonctionnement des organismes ; on appelle *gènes* les segments d'ADN sur lesquels ces instructions se retrouvent. De plus, un *locus* (loci au pluriel) désigne un emplacement précis sur une séquence d'ADN, nous parlerons donc du locus d'un gène afin de désigner son emplacement sur un chromosome.

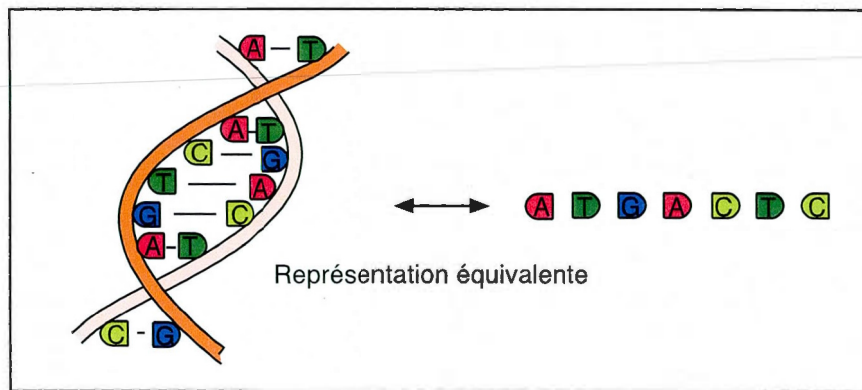


Figure 1.1 Deux représentations équivalentes d'une molécule d'ADN.

1.1.3 Les gènes

Il existe des milliers de gènes sur chaque chromosome et chacun de ceux-ci a une influence sur un ou plusieurs caractères héréditaires d'un organisme. L'influence qu'a un gène sur un organisme est définie par la version du gène, appelée *allèle*, que possède cet organisme. Par exemple, il existe trois allèles différents pour le gène ABO responsable du groupe

sanguin d'un individu : A, B et O. Il est important de noter qu'un individu diploïde possède deux copies de chaque gène et donc deux allèles codant pour le même caractère. En reprenant l'exemple des groupes sanguins pour des individus diploïdes, on obtient les six combinaisons d'allèles différentes représentées au tableau 1.1 et chacune de ces combinaisons est appelée *génotype* pour le gène ABO. Le génotype d'un individu diploïde est donc la composition allélique de tous ces gènes ; pour un gène précis, il peut être soit homozygote (deux allèles identiques), soit hétérozygote (2 allèles différents). Notons que le génotype d'un individu nous informe des deux allèles présents pour chacun de ses gènes, mais qu'il ne nous informe pas sur la manière dont ceux-ci sont répartis sur les deux chromosomes homologues. Lorsque cette répartition est connue, on ne parle plus de génotype, mais plutôt de *diplotype* ; celui-ci est composé de deux *haplotypes*, chacun représentant la composition allélique des gènes d'un des deux chromosomes homologues. La figure 1.2 illustre le génotype, le diplotype et les deux haplotypes d'un individu hétérozygote pour deux gènes donnés. Notons que tout au long de ce mémoire, nous supposons que la répartition des allèles sur les chromosomes homologues d'un individu est connue, nos données seront donc sous forme d'haplotypes plutôt que de génotypes.

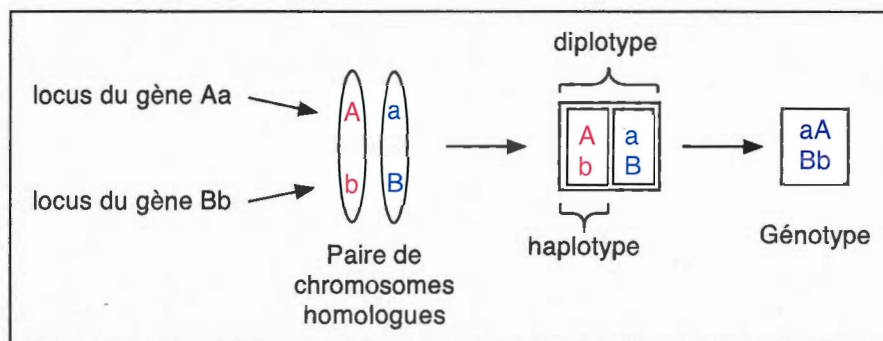


Figure 1.2 Représentation imagée du génotype et du diplotype pour les gènes Aa et Bb.

L'état du caractère, sur lequel un gène a une influence, est appelé *phénotype* ; il est observable et peut être de plusieurs natures (anatomique, morphologique, etc.). Dans

Génotypes	AA	AB	AO	BB	BO	OO
Phénotypes	A	AB	A	B	B	O

Tableau 1.1 Lien entre le génotype et le phénotype du gène ABO.

l'exemple du gène ABO, il y a quatre phénotypes possibles qui représentent les quatre différents groupes sanguins chez l'humain, soit A, B, AB et O. Le tableau 1.1 illustre le phénotype lié à chacun des six génotypes possibles. En observant ce tableau, on constate que l'allèle O s'exprime uniquement lorsqu'il y a présence de deux copies de cet allèle, on dit alors que l'allèle O est *récessif*. Pour leur part, les allèles A et B sont dit *dominants* puisqu'une seule copie de leur allèle est suffisante à leur expression dans le phénotype. Notons que le phénotype qui nous intéressera dans les chapitres subséquents sera d'être atteint ou non par une maladie génétique d'intérêt. Un individu atteint de la maladie sera considéré comme un *cas*, sinon il sera appelé un *témoin* ou un *contrôle*.

1.2 Transmission du bagage génétique chez l'humain

Le corps humain est constitué de plusieurs milliards de cellules, celles-ci peuvent être divisées en deux catégories en fonction du nombre de chromosomes que l'on retrouve dans leur noyau. Les cellules haploïdes ou reproductrices sont celles contenant un unique jeu de 23 chromosomes, tandis que les cellules diploïdes ou somatiques en contiennent deux, c'est-à-dire 23 paires de chromosomes homologues. Les cellules reproductrices de la femme sont les ovules et celles de l'homme les spermatozoïdes. Toutes les autres cellules de l'être humain sont somatiques. Lorsqu'il y a fécondation d'un ovule par un spermatozoïde, on obtient une cellule diploïde, appelée zygote, ayant hérité du bagage génétique de ses deux parents. En effet, chaque paire de chromosomes homologues de cette cellule est constituée d'un chromosome provenant de la mère et d'un chromosome provenant du père.

Les cellules reproductrices sont formées à partir de cellules somatiques lors d'une phase du cycle cellulaire appelée *méiose*. Au cours de ce processus, une cellule diploïde réplique

son ADN avant de se diviser deux fois, afin de former 4 cellules haploïdes, chacune formée de la moitié des chromosomes de la cellule mère. Il est intéressant de constater qu'une grande variation génétique découle de la méiose. En effet, il existe 2^{23} façon de transformer les deux jeux de 23 chromosomes d'une cellule en un seul, donc environ 8 millions de combinaisons haploïdes possibles, ce qui donne plus de 64 billions de combinaisons diploïdes. Ceci est donc une contribution non négligeable au fait que les descendants d'un individu ne sont pas identiques, à l'exception bien sûr des jumeaux identiques qui proviennent du même zygote. La section suivante présente deux autres sources majeures de variations génétiques qui seront d'un grand intérêt tout au long de ce mémoire.

1.3 Variabilité génétique

1.3.1 Sources de variation génétique

Il a été mentionné dans la section précédente que la méiose apporte une grande variation génétique par la sélection aléatoire du jeu de chromosome qui sera transmis à une cellule reproductrice. Cependant, un mécanisme se produisant lors de la méiose contribue aussi à la variation génétique, celui-ci est appelé *enjambement*. Un enjambement se produit lorsque deux chromosomes homologues échangent des segments avant que la cellule ne subisse sa première division. La figure 1.3 illustre la méiose d'une cellule possédant une unique paire de chromosome sur laquelle un enjambement s'est produit. On constate que la moitié des cellules produites possèdent une nouvelle combinaison génétique sur leur chromosome, celui-ci étant formé d'un mélange du matériel génétique des chromosomes parentaux; ce fait sera très important par la suite. Ainsi, les enjambements sont une grande source de brassage génétique.

On dit qu'il y a eu une *recombinaison* entre deux gènes (ou de façon générale entre deux loci) d'un même chromosome, lorsque le matériel génétique aux deux loci en question provient de deux chromosomes différents. Un événement de recombinaison entre deux gènes se produit donc lorsqu'il y a un nombre impair d'enjambements entre ceux-ci. La

figure 1.4 illustre la raison pour laquelle un nombre pair d'enjambements entre deux gènes ne cause pas de recombinaison. Notons que la probabilité qu'un double enjambement se produise entre deux gènes rapprochés est faible, puisqu'il y a en moyenne 2 ou 3 enjambements par chromosome par méiose.

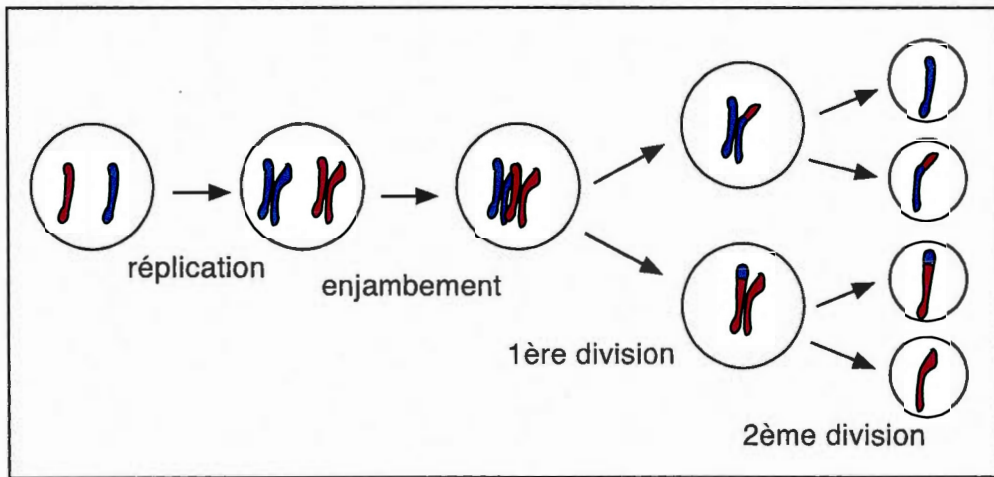


Figure 1.3 Illustration de la méiose d'une cellule diploïde possédant une unique paire de chromosomes homologues et où un enjambement s'est produit.

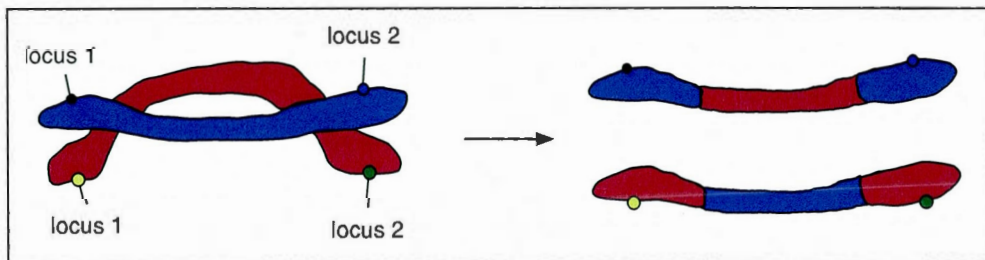


Figure 1.4 Illustration d'un double enjambement entre les gènes situés aux loci 1 et 2. Il n'y a pas de recombinaison entre ces gènes puisque le matériel génétique sur le chromosome bleu pour ces deux gènes (illustré par le cercle noir et le cercle mauve) est toujours situé sur le même chromosome suite au double enjambement. De la même façon, le matériel génétique du chromosome rouge pour ces deux gènes (illustré par un cercle jaune et un cercle vert) appartient lui aussi au même chromosome.

Une autre source de brassage génétique est la *mutagenèse*, c'est-à-dire l'apparition de mutations à l'intérieur d'un gène. Ces mutations peuvent être le résultat d'erreurs lors de la réplication ou de la recombinaison de l'ADN. Elles peuvent affecter la base azotée d'un ou de plusieurs nucléotides et ceci de trois façons différentes : certaines paires de bases peuvent avoir été substituées (substitution) ou même effacées (délétion), tandis que d'autres paires de bases peuvent avoir été ajoutées (insertion) dans une séquence d'ADN. Une mutation est dite silencieuse lorsqu'elle n'affecte pas l'allèle d'un gène ; une telle mutation n'est pas observable au point de vue phénotypique. De plus, une mutation n'affectant pas les capacités reproductrices de l'individu qui la porte est appelée mutation neutre. Pour sa part, une mutation affectant l'expression d'un gène peut avoir des effets néfastes sur la santé, une maladie causée par une telle mutation est appelée maladie génétique. Notons que l'objectif de ce mémoire est de développer une méthode permettant de trouver l'emplacement de telles mutations sur une séquence d'ADN. Il est important de mentionner que pour la suite de ce mémoire, nous allons uniquement considérer les mutations de type «substitution» affectant un unique nucléotide.

1.3.2 Distances et cartographie génétique

Il a été démontré en 1917 par Thomas Hunt Morgan et ses collaborateurs que la fréquence de recombinaison entre deux gènes est proportionnelle à la distance qui les sépare. En effet, plus deux gènes sont éloignés sur un chromosome, plus il y aura de chance pour qu'une recombinaison survienne entre ceux-ci. Cette découverte a permis de créer la *distance génétique* appelée centimorgan (cM), 1 cM étant équivalent à une probabilité d'enjambement de 1% entre deux gènes, et d'élaborer les premières cartes génétiques basées sur les recombinaisons.

Il existe aussi une unité de mesure pour la *distance physique* entre deux gènes : la paire de bases (bp, de l'anglais «base pair»); la distance correspond au nombre de nucléotides qui les séparent. Le nombre de nucléotides constituant l'ADN étant très élevé, il est courant d'employer la mégabase (Mb), correspondant à 1 000 000 de paires de bases. Chez l'humain, la relation approximative entre la distance génétique et la

distance physique est $1 \text{ cM} \approx 1 \text{ Mb}$.

Le terme cartographie génétique peut être utilisé dans deux situations complémentaires, toutes deux basées sur les variations observées entre les génomes d'individus. La première est celle où le terme cartographie génétique correspond au fait d'élaborer des cartes génétiques sur lesquelles nous retrouvons l'ordre des loci des gènes d'un chromosome ainsi que la distance génétique ou physique entre ceux-ci (voir figure 1.5 - situation 1). La deuxième situation est celle où le terme cartographie génétique correspond au fait d'identifier sur une séquence d'ADN le (ou les) gène(s) influençant un caractère héréditaire donné. Lorsque ce caractère héréditaire est d'avoir ou non une maladie génétique, la cartographie génétique peut aussi faire référence au fait de trouver l'emplacement à l'intérieur d'un gène de la mutation responsable de cette maladie (voir figure 1.5 - situation 2). Nous parlons de cartographie génétique fine lorsque la recherche de l'emplacement d'une telle mutation est effectuée sur une petite région du génome. Pour la suite de ce mémoire, le terme cartographie génétique référera au fait de trouver l'emplacement sur une séquence d'ADN d'une mutation responsable d'une maladie génétique.

1.3.3 Marqueurs génétiques

Le génome humain contient d'immenses portions de séquences d'ADN qui sont identiques dans la population; ces portions ne sont pas utilisées en cartographie génétique, car elles ne contiennent pas d'information sur les variations entre les individus. Il sera donc nécessaire d'utiliser des séquences d'ADN polymorphes (qui ne sont pas les mêmes chez tous les individus) afin d'obtenir des résultats optimaux. C'est pourquoi les marqueurs génétiques sont très employés en cartographie génétique. Un marqueur génétique est une séquence d'ADN située dans le génome qui présente des variations facilement détectables en laboratoire. Il existe plusieurs sortes de marqueurs génétiques; celui utilisé dans les méthodes présentées dans ce mémoire est le *polymorphisme nucléotide simple* (SNP, de l'anglais single-nucleotide polymorphism). Ce polymorphisme de l'ADN survient lorsque sur une courte séquence d'ADN, tous les individus d'une population possèdent la même

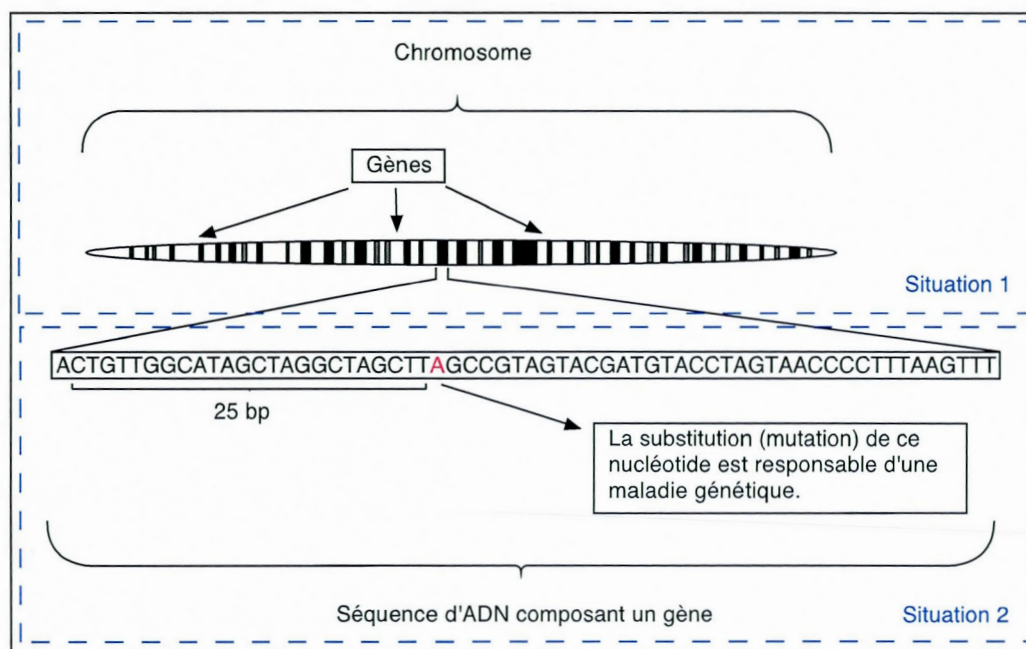


Figure 1.5 Exemple illustrant les deux situations dans lesquelles le terme cartographie génétique peut être employé. Dans le rectangle tracé en pointillé bleu représentant la situation 1, nous retrouvons l'illustration d'un chromosome sur lequel les rectangles noirs, gris foncés, gris pâles et blancs représentent les loci des gènes. Notons que les distances génétiques entre les gènes n'ont pas été représentées afin de ne pas surcharger l'illustration. Dans la deuxième situation, la séquence d'ADN d'un des gènes du chromosome est illustrée ; le nucléotide responsable d'une maladie génétique quelconque a été dessiné en rose. Ce nucléotide est situé à une distance physique de 25 bp du premier nucléotide contenu dans ce gène.

séquence à l'exception d'une paire de bases ; cette paire de bases est un SNP. De plus, bien que l'ADN soit formé d'une suite de 4 différents nucléotides, il existe en général seulement deux nucléotides possibles par SNP et donc par le fait même, uniquement deux allèles possibles. La façon courante de représenter les deux allèles d'un SNP est d'utiliser une notation binaire. L'allèle représentant la version originale d'un marqueur, c'est-à-dire l'allèle présent sur le marqueur avant que celui-ci ait subi une mutation est dit *primitif* et on lui assignera l'allèle «0» tandis que l'allèle représentant la version

mutée d'un marqueur, c'est-à-dire l'allèle présent sur le marqueur après que celui-ci ait subi une mutation, est dit *dérivé* (ou mutant) et on lui assignera l'allèle «1». La figure 1.6 illustre quatre séquences d'ADN provenant de différents individus où les nucléotides surlignés en jaune représentent des SNP. Nous y retrouvons de plus une représentation possible des séquences de SNP sous forme allélique.

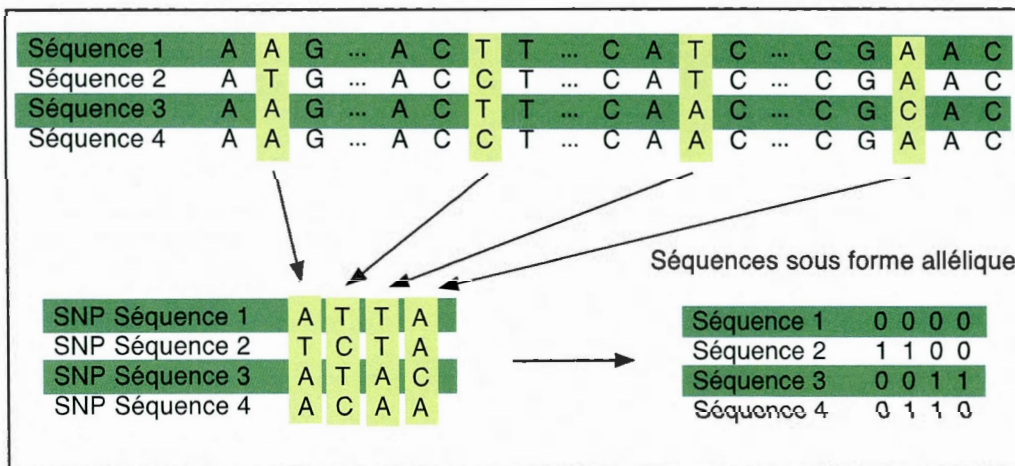


Figure 1.6 Exemple illustrant la conversion des données brutes en SNP sous forme allélique.

Les SNPs sont stables, abondants et leurs emplacements sont distribués uniformément dans le génome. Ils forment donc un repère génétique idéal et grandement utilisé dans les études génétiques actuelles. Ainsi, pour la suite de ce mémoire, nous ne considérerons plus des séquences d'ADN, mais bien des séquences de SNPs, que nous appellerons séquences génétiques. De plus, les échantillons de données considérés par les méthodes de cartographie génétique que nous présenterons dans ce mémoire seront formés de séquences génétiques provenant d'individus non-relés entre eux. L'être humain étant un être diploïde, chaque individu contribuera à un échantillon avec deux séquences génétiques provenant d'une certaine région de leur chromosome.

CHAPITRE II

THÉORIE DE LA COALESCENCE

Le modèle sous-jacent à la nouvelle méthode de cartographie génétique présentée dans ce mémoire est basé sur la relation généalogique qui existe entre les séquences génétiques provenant d'individus non-reliés entre eux. Cette relation est malheureusement inconnue pour la plupart des échantillons de séquences génétiques ; l'approche employée afin de pallier à ce problème est de générer des généalogies plausibles et compatibles avec les séquences génétiques d'un échantillon donné. La génération de telles généalogies sera faite à l'aide d'une adaptation du graphe de recombinaison ancestral. Ce chapitre a donc pour but de présenter le graphe de recombinaison ancestral, mais pour ce faire, nous devons tout d'abord introduire le modèle de Wright-Fisher et le processus de coalescence.

2.1 Le modèle de Wright-Fisher

Fisher (1930) et Wright (1931) ont introduit un modèle de base servant à modéliser l'évolution génétique d'une population, c'est-à-dire la façon dont les séquences génétiques se transmettent d'une génération à une autre. Le modèle est très simple et il est basé sur les suppositions suivantes :

- les générations sont discrètes et ne se chevauchent pas : ceci signifie que tous les individus d'une génération ont la même durée de vie, ils naissent et meurent donc tous simultanément ;
- la taille de la population est constante et finie : par convention, on pose cette taille

- égale à $2N$;
- les individus sont haploïdes : un individu est équivalent à une séquence génétique ;
- il n'y a pas de sélection naturelle ni de structure particulière dans la population : ceci signifie que tous les individus d'une même génération ont la même probabilité de se reproduire ;
- les séquences ne peuvent pas recombinaison.

La simulation de l'évolution d'une population de séquences génétiques avec le modèle de Wright-Fisher est très simple. On débute avec une population génétique initiale et par la suite, chaque nouvelle génération est formée d'un échantillon aléatoire avec remise des séquences génétiques de la génération précédente. La figure 2.1(a) illustre ce phénomène pour une population de taille $2N = 10$, chaque séquence génétique y est représentée par un cercle orange. Une fois l'évolution d'une population générée, il est facile d'obtenir la généalogie d'un échantillon de séquences appartenant à la dernière génération simulée. En effet, il suffit de remonter le temps en suivant les lignées des séquences de l'échantillon et ce, jusqu'à ce qu'il n'y ait qu'une seule lignée restante ; celle-ci est appelée l'*ancêtre commun le plus récent*, que l'on notera MRCA de l'anglais «most recent common ancestor». La figure 2.1(b) illustre la généalogie d'un échantillon contenant trois séquences. En observant attentivement cette représentation, on constate que les seuls éléments qui nous apportent de l'information sur la généalogie sont la connaissance des séquences qui ont trouvé un ancêtre commun et le nombre de générations qu'il a fallu pour que de tels événements se produisent. Ceci s'illustre par le fait qu'on ne perd aucune information sur la généalogie de notre échantillon en la représentant à l'aide de la figure 2.2. La prochaine section est consacrée à la présentation du processus de coalescence qui utilise l'observation précédente afin de générer des généalogies du présent vers le passé.

2.2 Le processus de coalescence

Définissons tout d'abord le terme *coalescence* ; ce terme est utilisé lorsque deux séquences trouvent un ancêtre commun, on dit alors que les deux séquences ont coalescé ou qu'il y a eu une coalescence entre ces deux séquences. Le processus de coalescence, introduit

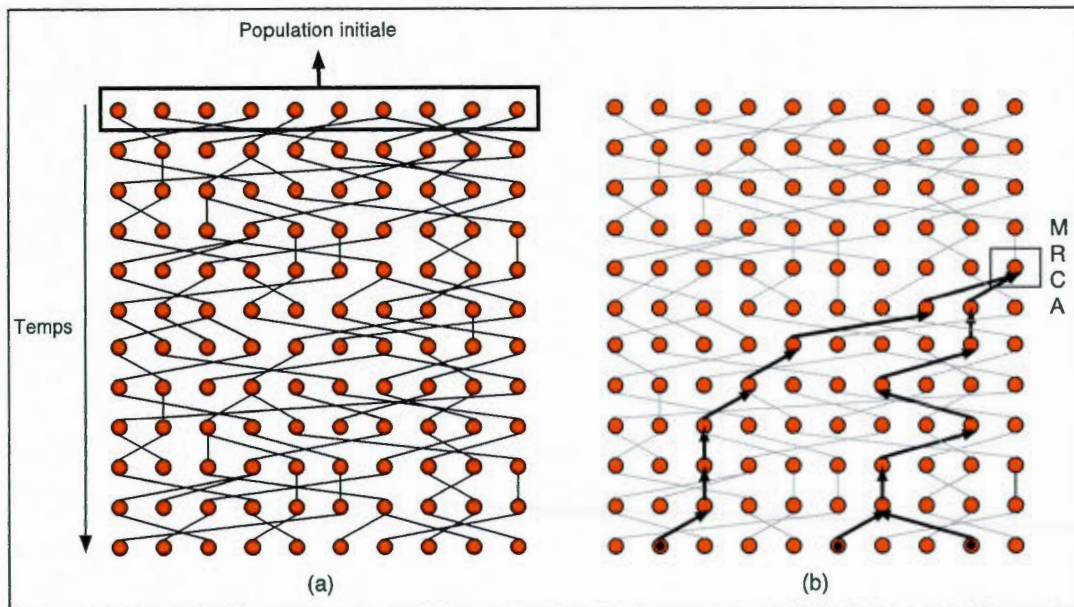


Figure 2.1 Exemple de l'évolution d'une population sous le modèle Wright-Fisher. La figure (a) représente la simulation de 12 générations d'une population de taille 10. La génération la plus récente se trouve au bas de la figure et les traits noirs entre les générations indiquent quelles séquences ont été choisies aléatoirement afin former la génération suivante. Les flèches noires de la figure (b) représentent la généalogie de l'échantillon formé des trois séquences marquées d'un point noir.

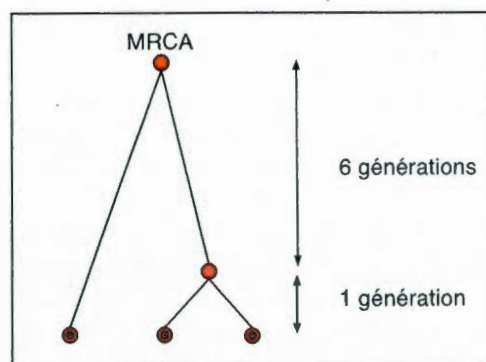


Figure 2.2 Représentation de la généalogie illustrée à la figure 2.1(b), une fois celle-ci isolée de la population.

par Kingman (Kingman, 1982), est un processus stochastique à temps continu décrivant la généalogie des séquences génétiques provenant d'un échantillon de taille fixe n tiré d'une population de taille $2N$. Les états possibles de ce processus stochastique sont tous les sous-ensembles possibles de l'ensemble contenant les séquences génétiques de notre échantillon ; un état correspond en fait aux séquences présentes à une génération d'un généalogie. Le passage d'un état à un autre s'effectue lorsqu'un événement de coalescence survient, c'est-à-dire lorsque deux séquences coalescent. Il a été démontré que c'est un «processus limite» pour le modèle de Wright-Fisher ; en effet, lorsque N tend vers l'infini et que n est beaucoup plus petit que N , les deux modèles sont équivalents. Comme il a été mentionné à la section précédente, ce processus est utilisé afin de générer la généalogie d'un échantillon de séquences génétiques et ce, du présent vers le passé. Pour ce faire, il suffit de générer des événements de coalescence et des temps auxquels ces événements se produisent.

2.2.1 Coalescence à temps discret

Afin de faciliter la description du processus de coalescence, étudions tout d'abord les événements de coalescence lorsque le temps est discret, c'est-à-dire qu'une unité de temps est égale à une génération. Posons C_{ij}^k l'événement où deux séquences i et j coalescent k générations dans le passé, la probabilité d'un tel événement est :

$$P[C_{ij}^k] = \frac{1}{2N} \cdot \left(1 - \frac{1}{2N}\right)^{k-1}, \quad (2.1)$$

où $1/2N$ est la probabilité que les séquences i et j coalescent à la génération précédente. L'équation 2.1 décrit donc le fait que les séquences i et j n'ont pas trouvé d'ancêtre commun pendant $k-1$ générations et qu'elles ont finalement coalescé il y a k générations. On déduit de cette équation que le temps requis pour que deux séquences trouvent un ancêtre commun suit une loi géométrique de paramètre $1/2N$. Il faudra donc en moyenne $2N$ générations pour que deux séquences particulières trouvent un ancêtre commun.

En utilisant un raisonnement similaire, on obtient que la probabilité que n séquences

aient des ancêtres distincts dans la génération précédente, noté $P(n)$, est :

$$\begin{aligned}
 P(n) &= \left(1 - \frac{1}{2N}\right) \left(1 - \frac{2}{2N}\right) \dots \left(1 - \frac{n-1}{2N}\right) \\
 &= 1 - \frac{1}{2N} - \frac{2}{2N} - \frac{3}{2N} - \dots - \frac{n-1}{2N} + O(1/N^2) \\
 &= 1 - \frac{1}{2N} \cdot \left(1 + 2 + 3 + \dots + (n-1)\right) + O(1/N^2) \\
 &= 1 - \frac{1}{2N} \cdot \left(\frac{(n-1)n}{2}\right) + O(1/N^2) \\
 &= 1 - \frac{1}{2N} \binom{n}{2} + O(1/N^2),
 \end{aligned}$$

où $O(1/N^2)$ représente tous les termes divisés par une puissance de N plus grande ou égale à 2, c'est-à-dire les termes qui correspondent au fait que plus de deux séquences coalescent en même temps. Ce terme est négligeable puisque nous avons supposé que $n \ll N$; la probabilité qu'il n'y ait pas d'événement de coalescence dans une génération donnée est donc de $1 - \binom{n}{2} \frac{1}{2N}$. On en déduit que la probabilité que le temps d'attente T_c^n avant que deux séquences quelconques parmi un échantillon de n séquences trouvent un ancêtre commun soit de k générations est :

$$P(T_c^n = k) \approx \left[1 - \binom{n}{2} \frac{1}{2N}\right]^{k-1} \binom{n}{2} \frac{1}{2N}, \quad (2.2)$$

et que T_c^n suit approximativement une loi géométrique de paramètre $\binom{n}{2} \frac{1}{2N}$.

2.2.2 Coalescence à temps continu

L'approche classique afin de modéliser le temps sur une échelle continue est de considérer qu'une unité de temps est égale à $2N$ générations. On peut ainsi approximer la loi géométrique de l'équation 2.2 par une loi exponentielle : $T_c^n \sim \exp(\binom{n}{2})$. Maintenant que l'on connaît la loi du temps d'attente avant une coalescence, il est facile de simuler une généalogie pour un échantillon de taille n : il suffit de simuler un temps d'attente $t \sim \exp(\binom{n}{2})$ et ensuite de faire coalescer deux lignées choisies aléatoirement parmi les n présentes. Après cette étape, il y aura une lignée en moins, on peut donc répéter cette procédure $(n-1)$ fois afin d'obtenir à la fin une unique lignée qui correspondra au MRCA de notre échantillon initial de taille n .

2.3 Processus de coalescence avec mutation

Il a été mentionné au chapitre précédent que l'apparition de mutation est une source non-négligeable de diversité génétique dans une population, il est donc intéressant d'incorporer des événements de mutation au processus de coalescence. Cette sous-section est donc consacrée à l'étude d'un modèle pour l'ajout des mutations.

Il est important de se rappeler tout d'abord qu'une des suppositions énoncées à la section 2.1 stipule qu'on considère seulement des populations où la sélection naturelle est absente. Ceci signifie que les événements de mutation n'ont pas d'impact sur la succession des générations, on peut donc superposer les mutations sur la généalogie une fois celle-ci simulée. Il existe plusieurs modèles mathématiques afin de traiter de la superposition des mutations sur une généalogie ; celui utilisé tout au long de ce mémoire est le modèle des sites infinis. Dans ce modèle, les séquences génétiques sont considérées comme une suite infinie de loci, et le locus du marqueur qui devra muter est choisi aléatoirement sur la séquence. Il en résulte donc qu'il ne peut pas y avoir plus d'un événement de mutation à un même locus. Suite à la superposition de mutations sur une généalogie à l'aide de ce modèle, les séquences résultantes sont habituellement présentées uniquement par les sites où il y a eu une mutation ; les autres loci ne sont pas informatifs puisque toutes les séquences y partagent les mêmes allèles. Une façon classique de représenter de telles données est d'utiliser le même choix de deux allèles pour chaque site : «0» si l'allèle à ce site descend directement du MRCA, c'est-à-dire lorsque l'allèle est primitif et «1» si l'allèle a subi une mutation au cours de son histoire, c'est-à-dire lorsque l'allèle est dérivé. La conversion des séquences d'un échantillon en représentation binaire est illustrée à la figure 2.4(b) que l'on retrouve à la toute fin de ce chapitre. Notons que le modèle des sites infinis est très approprié pour modéliser les mutations apparaissant sur une généalogie de séquences génétiques composées de SNPs. En effet, les SNPs sont des marqueurs qui ont un taux de mutation très faible (on parle ici d'un taux de mutation de l'ordre de 10^{-8} par séquence par génération), la probabilité qu'il y ait eu plus d'une mutation à un même locus au cours de l'histoire est donc quasiment nulle.

Étudions maintenant la façon d'incorporer le modèle de mutation choisi au processus de coalescence. En posant μ la probabilité d'un événement de mutation par séquence par génération, on obtient que la probabilité que le temps d'attente T_m avant qu'un événement de mutation survienne sur une lignée soit égal à k générations est :

$$P(T_m = k) = \mu(1 - \mu)^{(k-1)}, \quad (2.3)$$

et que T_m suit une loi géométrique de paramètre μ . Encore une fois, en se plaçant sur une échelle à temps continu et en supposant que N est très grand, on peut approximer T_m par une loi exponentielle : $T_m \sim \exp(2N\mu)$. Il est courant, afin de simplifier certaines équations, de poser $\theta = 4N\mu$ et d'ainsi obtenir : $T_m \sim \exp(\theta/2)$. Notons que ceci est équivalent mathématiquement à ce que le nombre de mutation apparaissant sur une branche de longueur t d'un généalogie, noté M_t , ait une distribution de Poisson d'intensité $t\theta/2$. On obtient donc la probabilité suivante :

$$P(M_t = j) = \frac{(t\theta)^j}{j!2^j} \exp(-t\theta/2).$$

Nous pouvons remarquer de plus que le processus permettant d'ajouter les mutations sur une branche d'une généalogie avec l'équation précédente est en fait un processus de Poisson. Ceci a pour effet que sachant le nombre de mutations qui sont apparues sur une branche, les temps auxquels ces mutations sont survenues suivent une distribution uniforme ; cette propriété sera utilisée au chapitre 4.

Considérons maintenant n lignées indépendantes, on obtient alors que le temps d'attente T_m^n avant qu'un événement de mutation survienne sur une de ces n lignées suit une loi exponentielle de paramètre égal à la somme des taux des n lois exponentielles considérées. De plus, les événements de coalescence et de mutation arrivant de façon indépendante, on obtient que le temps d'attente avant qu'un de ces deux événements survienne suit une loi exponentielle de paramètre :

$$\binom{n}{2} + \frac{n\theta}{2} = \frac{n(n-1+\theta)}{2}.$$

Grâce à cette dernière information, il est facile de simuler une généalogie d'une façon similaire à celle décrite dans la section 2.2.2, mais cette fois-ci en y ajoutant des

événements de mutation. En effet, lorsque la taille de l'échantillon pour une génération t est de k , on simule la génération $t + 1$ de la façon suivante :

- simuler le temps d'attente T avant le prochain événement, avec $T \sim \exp(\frac{k(k-1+\theta)}{2})$;
- le prochain événement est une coalescence avec probabilité $\frac{k-1}{k-1+\theta}$ et un événement de mutation avec probabilité $\frac{\theta}{k-1+\theta}$;
- si l'événement est une coalescence, choisir aléatoirement deux lignées et les faire coalescer. Sinon, apposer une mutation sur une lignée choisie aléatoirement et choisir le locus du marqueur qui devra muter.

2.4 Le graphe de recombinaison ancestral

Les recombinaisons génétiques, tel qu'indiqué à la section 1.3, sont des événements fréquents qui contribuent à la diversité génétique. Chez l'humain, tous les chromosomes peuvent subir des recombinaisons à l'exception du chromosome Y de l'homme ; il est donc important de considérer ces événements lorsque l'on modélise la généalogie d'un échantillon de séquences génétiques. Cette section est donc dédiée à la théorie de la coalescence avec recombinaison.

Tout d'abord, il est important de rappeler que les recombinaisons ne peuvent arriver que dans des cellules diploïdes ; nous devons donc adapter notre modèle afin qu'il puisse considérer des individus diploïdes. Comme nous l'avons mentionné à la section 1.1.3, nous supposons que les deux haplotypes de chaque individu diploïde constituant un échantillon sont connus. Ainsi, afin de pouvoir incorporer des événements de recombinaison dans notre modèle, il suffit de considérer une population de N individus diploïdes, car cette population est équivalente à une population de $2N$ individus haploïdes et peut donc être modélisée à l'aide du modèle présenté précédemment.

En 1983, Hudson (Hudson, 1983) a intégré les recombinaisons dans le processus de coalescence. Son modèle pour les recombinaisons est plutôt simple. En effet, afin de simuler la recombinaison de deux séquences, il suffit de choisir aléatoirement un point de recombinaison et ensuite de former la séquence résultante telle que son matériel génétique à la

droite du point de recombinaison provienne d'une des deux séquences parentales et celui à la gauche provienne de l'autre séquence parentale. Ce phénomène est illustré à la figure 2.3(a). Cette modélisation est adaptée au cas où on simule du passé au présent, or le processus de coalescence permet de simuler des généalogies rétrospectivement. On doit donc adapter le modèle à cette condition. Un événement de recombinaison, observé du présent au passé, se produit lorsqu'une séquence sépare son matériel génétique afin de former deux nouvelles séquences. Afin de simuler ce phénomène, il suffit encore une fois de choisir aléatoirement un point de recombinaison et ensuite de créer deux nouvelles séquences telles que le matériel génétique à la droite du point de recombinaison soit transmis à une des séquences et celui à la gauche soit transmis à l'autre séquence. En observant ce phénomène sur la figure 2.3(b), on constate qu'une recombinaison a pour effet de diviser une séquence en deux, et correspond donc à l'effet opposé d'un événement de coalescence. De plus, on observe que du matériel génétique inconnu (section avec les points d'interrogation) a été introduit dans notre généalogie; ce matériel est appelé *matériel non-ancestral*. Le matériel non ancestral est constitué de matériel génétique qui n'est pas transmis aux séquences génétiques de notre échantillon, et ce, à l'opposé des marqueurs primitifs et dérivés, ces deux derniers types de marqueurs forment donc le matériel appelé *matériel ancestral*. La présence de matériel non-ancestral dans une généalogie n'est pas problématique, puisque le fait qu'il ne se retrouve pas dans les séquences de notre échantillon le rend non-informatif.

Analysons maintenant la façon d'introduire des événements de recombinaison dans le processus de coalescence. En posant r la probabilité d'un événement de recombinaison par séquence par génération, on obtient que la probabilité que le temps T_r avant qu'un événement de recombinaison survienne sur une lignée soit égal à k générations est :

$$P(T_r = k) = r(1 - r)^{(k-1)},$$

et que T_r suit une loi géométrique de paramètre r . Encore une fois, en se plaçant sur une échelle à temps continu et en supposant que N est très grand, on peut approximer T_r par une loi exponentielle : $T_r \sim \exp(2Nr)$. Analogiquement au taux de mutation de la population θ , on définit ρ le taux de recombinaison de la population comme $\rho = 4Nr$

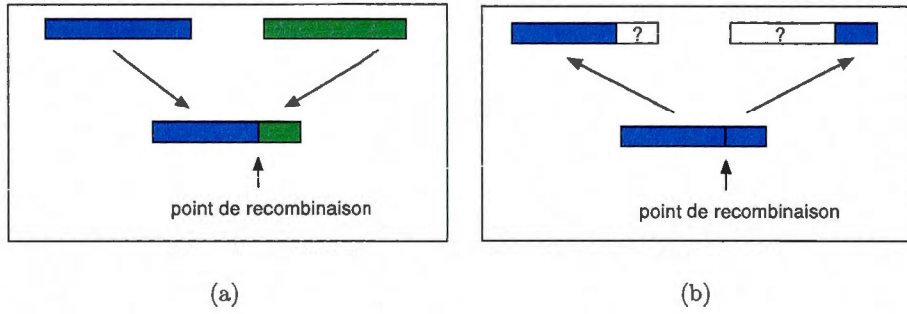


Figure 2.3 Illustration d'un événement de recombinaison. La figure (a) représente cet événement du passé vers le présent et la figure (b) l'illustre du présent vers le passé. Les points d'interrogations sur cette dernière figure soulignent le fait que ces portions des séquences parentales sont inconnues lorsqu'on simule rétrospectivement.

et ainsi : $T_r \sim \exp(\rho/2)$.

Puisque les événements de recombinaison sont indépendants de ceux de coalescence et de mutation, d'une façon analogue à la section 2.3, on obtient que le temps d'attente avant le prochain événement, lorsqu'il a n lignées présentes à une génération, suit une loi exponentielle de paramètre :

$$\binom{n}{2} + \frac{n\theta}{2} + \frac{n\rho}{2} = \frac{n(n-1+\theta+\rho)}{2}. \quad (2.4)$$

L'algorithme permettant de générer une étape de la généalogie d'un échantillon présenté à la section 2.3 peut être adapté afin d'inclure des événements de recombinaison. On obtient alors l'algorithme suivant pour la simulation d'une généalogie complète, où k représente le nombre de séquences présentes dans une génération et n la taille de notre échantillon initial :

1. poser $k = n$;
2. tant que $k > 1$, faire :
 - (a) simuler le temps T avant le prochain événement, avec $T \sim \exp(\frac{k(k-1+\theta+\rho)}{2})$;
 - (b) le prochain événement est une coalescence avec probabilité $\frac{k-1}{k-1+\theta+\rho}$, une mutation avec probabilité $\frac{\theta}{k-1+\theta+\rho}$ et une recombinaison avec probabilité $\frac{\rho}{k-1+\theta+\rho}$;

- (c) – si l'événement est une coalescence, choisir aléatoirement deux séquences et les faire coalescer. Poser k égal à $k - 1$;
- si l'événement est une mutation, apposer une mutation sur une lignée choisie aléatoirement et choisir le locus du marqueur qui devra muter ;
- si l'événement est une recombinaison, choisir aléatoirement une séquence et un point de recombinaison sur celle-ci et créer deux nouvelles séquences telles qu'illustré à la figure 2.3(b). Poser k égal à $k + 1$;

La figure 2.4(a) représente un exemple de généalogie simulée à l'aide de l'algorithme précédent. On remarque que l'ajout des événements de recombinaison dans la généalogie fait en sorte que celle-ci ne se représente plus par un arbre, mais bien par un graphe ; ce graphe a été nommé *graphe de recombinaison ancestral* (ARG) par Hudson (1991). De plus, nous pouvons remarquer qu'en plus de nous permettre de générer une généalogie, cet algorithme nous permet de générer un échantillon de séquences génétiques. En effet, en observant la façon dont les mutations (cercles de couleur) se transmettent dans la généalogie, nous obtenons l'échantillon de séquences génétiques représenté à la figure 2.4(b).

Notons que la convergence de l'algorithme précédent est assurée, c'est-à-dire que lorsque l'on simule une généalogie, on est certain d'atteindre un MRCA. En effet, lorsque l'on ne considère que des événements de coalescence et de recombinaison, le nombre de séquences présentes à chaque génération d'un graphe peut être modélisé par un processus de naissance et de mort. Le taux de coalescence étant quadratique ($n(n+1)/2$) et celui de recombinaison étant linéaire ($n\rho/2$), les événements de coalescence seront prédominants et ceci nous garantit une convergence de l'algorithme.

Il est maintenant important de faire quelques remarques utiles à la compréhension des prochains chapitres. Nous venons de présenter un algorithme basé sur le processus de coalescence avec recombinaison permettant de générer rétrospectivement des graphes de recombinaison ancestraux et d'obtenir des échantillons de séquences génétiques. Les ARGs obtenus à l'aide de cet algorithme n'ont cependant pas été générés à partir

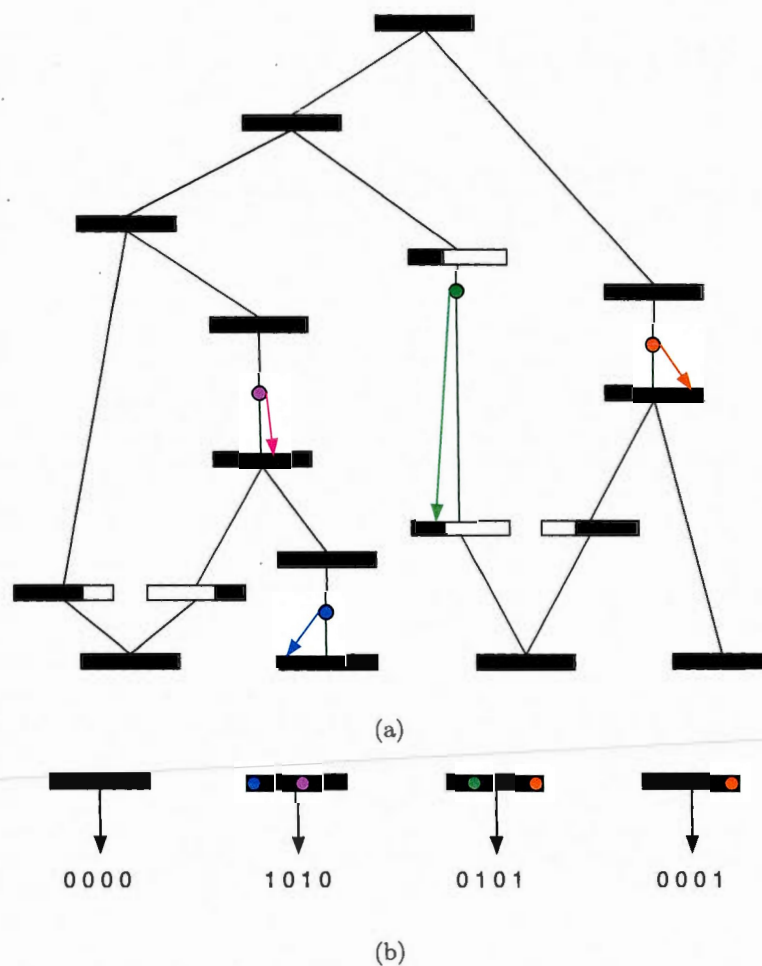


Figure 2.4 Illustration d'un graphe de recombinaison ancestral pour un échantillon de 4 séquences. La figure (a) représente un ARG où les flèches de couleur indiquent les loci où les mutations surviennent, et les portions de séquences blanches indiquent la présence de matériel non-ancestral. La figure (b) illustre l'échantillon obtenu suite à cette simulation et la façon de représenter les séquences obtenues sous forme binaire grâce au modèle de sites infinis pour les mutations. Rappelons que seuls les marqueurs ayant mutés au cours de leur histoire sont représentés sous forme binaire, ainsi, puisqu'il y a eu 4 événements de mutation dans cette généalogie, les séquences génétiques sont formées par une suite de 4 marqueurs.

d'un échantillon de séquences génétiques, comme nous le voudrions. C'est pourquoi nous devons utiliser une adaptation de l'algorithme précédent afin de générer de tels graphes, mais cette fois-ci à partir d'un échantillon de séquences génétiques donné; cette adaptation sera présentée en détail dans le prochain chapitre. Afin d'en faciliter la compréhension, nous allons énoncer les principales caractéristiques d'une généalogie générée à partir d'un échantillon de séquences génétiques. Pour ce faire, nous allons analyser la figure 2.5, qui représente le même ARG qu'à la figure 2.4(a), mais cette fois-ci généré à partir de l'échantillon illustré à la figure 2.4(b). On distingue dans cette généalogie quatre différents événements permettant la transition du présent au passé d'une génération à une autre. Un exemple de chacun de ces événements est illustré sur la figure par une lettre encerclée. L'événement A correspond à une coalescence entre deux séquences identiques. La lettre B indique un événement de coalescence entre deux séquences qui ne sont pas identiques, mais qui diffèrent uniquement par des portions contenant du matériel non ancestral. Un événement de mutation est illustré en C; on remarque que la mutation (cercle orange) n'est plus ajoutée sur la lignée, mais bien retirée. De plus, le modèle de mutation des sites infinis nous assure qu'il y a au plus une mutation par locus et ceci a pour conséquence qu'une mutation ne peut être retirée que lorsqu'elle apparaît sur une unique séquence de la génération considérée. Finalement, la lettre D illustre un événement de recombinaison dont le point de recombinaison est situé entre le 3^e et le 4^e marqueur. Notons de plus que l'ancêtre commun le plus récent est toujours formé de marqueurs ancestraux et primitifs; il est donc connu. Ceci est une conséquence du fait que nous l'utilisons comme séquence de référence afin de distinguer les allèles dérivés des allèles primitifs et afin de définir le matériel non-ancestral.

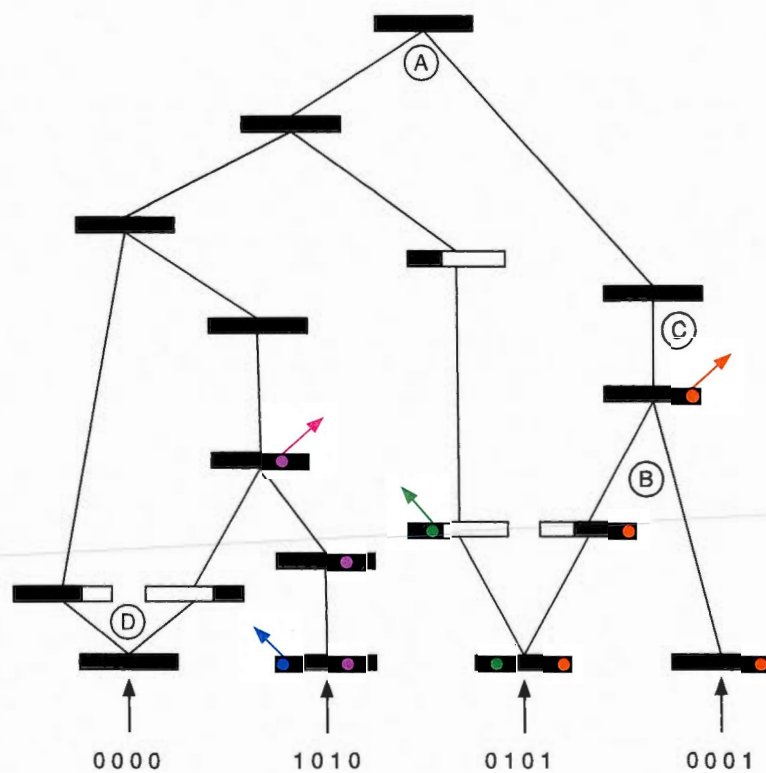


Figure 2.5 Illustration d'un ARG généré à partir d'un échantillon donné. La différence entre cet ARG et celui représenté à la figure 2.4(a) est au niveau des événements de mutation. En effet, lors d'un événement de mutation dans cet ARG, plutôt que d'ajouter une mutation sur une lignée comme nous le faisons précédemment, nous la retirons.

CHAPITRE III

MÉTHODES DE CARTOGRAPHIE GÉNÉTIQUE FINE

Il existe à ce jour plusieurs méthodes de cartographie génétique fine. Parmi celles-ci, trois nous intéressent particulièrement puisque la nouvelle méthode présentée dans ce mémoire en est inspirée. Ce chapitre est donc consacré à la présentation de ces trois méthodes, l'emphase étant mise sur les éléments qui ont servi d'inspiration à l'élaboration de la méthode DMap.

3.1 Éléments communs aux trois méthodes

Dans les trois méthodes de cartographie que nous allons présenter dans ce chapitre, les données utilisées sont de la même forme. En effet, tous les échantillons de données sont composés de courtes séquences génétiques ; rappelons que dans ce mémoire, nous appelons séquence génétique une suite de marqueurs de type polymorphisme nucléotide simple (SNP). De plus, chaque séquence provient d'un individu dont on connaît le phénotype pour la maladie génétique qui nous intéresse, c'est-à-dire que l'on sait s'il est un cas ou un témoin. Nous supposerons de plus que la maladie génétique considérée est influencée par une unique mutation, notée TIM de l'anglais «trait influencing mutation», et que cette mutation est apparue sur un marqueur appartenant à la portion du génome couverte par les séquences de notre échantillon. Cette dernière supposition est réaliste, car plusieurs méthodes de cartographie permettent d'estimer la région d'un chromosome sur laquelle un TIM est apparu. Le lecteur intéressé à de telles méthodes peut se référer aux articles de Lander et Schork (1994), de Olson *et al.* (1999) et de

Elston (2000).

3.2 Méthode MapARG

La méthode de cartographie génétique fine MapARG fut développée en 2002 (Larribe, Lessard et Schork, 2002) puis améliorée en 2008 (Larribe et Lessard, 2008). La nouvelle méthode présentée dans ce mémoire étant principalement une adaptation de celle-ci, les sections suivantes en contiennent une description détaillée.

3.2.1 Le modèle

Chaque séquence d'un échantillon a une longueur de r mégabase(s) et est formée de L marqueurs. Un seul marqueur parmi les L possède un allèle et une position inconnus : ce marqueur est celui correspondant au TIM. L'allèle de ce marqueur étant inconnu pour une séquence donnée, il devra être inféré à l'aide du phénotype (cas ou témoin) de la séquence en question et de d'autres informations éventuellement ; la méthode d'inférence ainsi que les difficultés liées à celle-ci sont décrites à la section 3.2.5.

On suppose que le TIM est situé entre le premier et le dernier marqueur connu d'une séquence, et on pose r_T la distance entre le TIM et le premier marqueur. L'objectif de la méthode est d'estimer à l'aide d'un estimateur à maximum de vraisemblance le paramètre r_T . Pour ce faire, la vraisemblance $L(r_T)$ sera calculée pour plusieurs valeurs candidates de r_T et l'estimation du paramètre r_T sera simplement la valeur candidate dont la vraisemblance est la plus élevée ; les valeurs candidates utilisées sont les distances entre le premier marqueur de la séquence et les milieux des $L - 2$ intervalles formés par les $L - 1$ marqueurs connus et elles sont notées π_1, \dots, π_{L-2} . La figure 3.1 illustre un échantillon contenant 3 séquences, chacune formée de $L - 1 = 4$ marqueurs connus, où le TIM a été inféré au milieu du deuxième intervalle.

En dénotant $H_0|r_T$ l'ensemble contenant les séquences génétiques de notre échantillon sur lesquelles le TIM a été inféré à une distance r_T du premier marqueur de notre

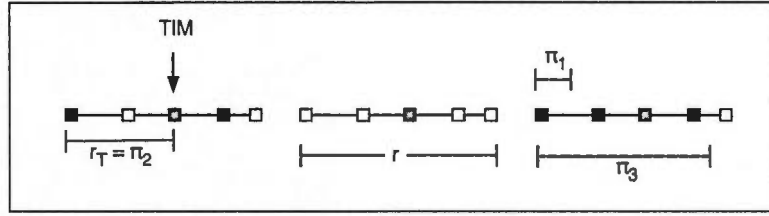


Figure 3.1 Illustration d'un échantillon contenant 3 séquences de $L - 1 = 4$ marqueurs connus sur lesquelles on a introduit le TIM dans le deuxième intervalle. Nous avons de plus illustré les distances r , π_1 , π_2 et π_3 . Un carré blanc représente un marqueur possédant un allèle primitif ou non mutant («0» en notation binaire), un carré noir représente un marqueur possédant un allèle dérivé ou mutant («1» en notation binaire) et un carré marbré blanc et noir représente un marqueur dont l'allèle est inconnu.

échantillon de départ, on obtient la fonction de vraisemblance suivante :

$$L(r_T) = P(H_0|r_T) = \int P(H_0|G, r_T) \cdot P(G|r_T) dG, \quad (3.1)$$

où l'intégrale est calculée sur l'espace des graphes de recombinaison ancestraux; G représente donc un ARG. Notons que nous utilisons une intégrale plutôt qu'une somme, car les branches d'un graphe de recombinaison ancestral sont distribuées selon une loi continue. De plus, $P(H_0|G, r_T)$ est une fonction indicatrice telle que :

$$P(H_0|G, r_T) = \begin{cases} 1 & \text{si l'échantillon obtenu par } G \text{ est le même que } H_0|r_T, \\ 0 & \text{sinon.} \end{cases}$$

Notons que lorsque $P(H_0|G, r_T) = 1$, la généalogie G est dite *consistante* avec l'ensemble de données $H_0|r_T$. Il est impossible d'évaluer exactement l'intégrale de l'équation 3.1, car l'espace des ARG est beaucoup trop vaste. Il est cependant possible de l'estimer à l'aide d'une approximation de Monte Carlo et on obtient alors l'équation suivante :

$$L(r_T) = P(H_0|r_T) = \int P(H_0|G, r_T) \cdot P(G|r_T) dG \approx \frac{1}{M} \sum_{i=1}^M P(H_0|G^{(i)}, r_T), \quad (3.2)$$

où $G^{(1)}, G^{(2)}, \dots, G^{(M)} \sim P(G|r_T)$.

Malheureusement, cette approximation est inefficace puisque la probabilité de générer un ARG consistant avec nos données est très faible, donc la plupart des termes contenus dans la somme de l'équation 3.2 seront nuls. Afin de pallier à l'inefficacité de l'approximation précédente, on utilise une méthode d'échantillonnage pondéré avec une distribution proposée notée $Q(G)$. On obtient alors :

$$\begin{aligned} L(r_T) &= \int \frac{P(H_0|G, r_T) \cdot P(G|r_T)}{Q(G)} \cdot Q(G) dG \\ &\approx \frac{1}{M} \sum_{i=1}^M \frac{P(H_0|G^{(i)}, r_T) \cdot P(G^{(i)}|r_T)}{Q(G^{(i)})} \end{aligned} \quad (3.3)$$

$$= \frac{1}{M} \sum_{i=1}^M \frac{P(G^{(i)}|r_T)}{Q(G^{(i)})}, \quad (3.4)$$

où les ARGs $G^{(i)}$, $i = 1, \dots, n$ sont simulés selon la fonction de densité $Q(G)$.

Notons que les équations 3.3 et 3.4 sont égales uniquement lorsque l'on suppose que les généalogies générées de la densité $Q(G)$ sont consistantes avec les données. Une façon d'obtenir de telles généalogies est de ~~les générer à partir de l'échantillon $H_0|r_T$ à l'aide~~ d'une adaptation de l'algorithme présenté à la section 2.4. Deux distributions proposées permettant de générer des généalogies à partir de l'échantillon $H_0|r_T$ seront présentées en détail dans la section 3.2.3.

3.2.2 La probabilité d'un graphe de recombinaison ancestral

Nous allons maintenant étudier la façon de calculer $P(G|r_T) \equiv P_{r_T}(G)$, où G est un ARG consistant avec nos données H_0 .

Posons H_τ l'ensemble contenant les séquences génétiques présentes à la génération τ du graphe G , avec $\tau = 0, \dots, \tau^*$; H_0 contient les séquences de notre échantillon et H_{τ^*} contient une unique séquence supposée connue : le MRCA. Le graphe de recombinaison ancestral G peut donc se définir du présent au passé par la suite des états : $H_0, H_1, \dots, H_{\tau^*}$; le passage d'un état à un autre se produit lorsqu'il y a un événement

de coalescence, de mutation ou de recombinaison. De plus, une génération k d'un graphe de recombinaison ancestral dépend du passé uniquement par la génération $(k - 1)$ la précédant. Il est alors possible de calculer la probabilité d'une généalogie à l'aide d'une récurrence sur ses états :

$$\begin{aligned}
 P_{r_T}(G) &= P_{r_T}(H_0, H_1, \dots, H_{\tau^*}) \\
 &= P_{r_T}(H_0|H_1) \cdot P_{r_T}(H_1, H_2, \dots, H_{\tau^*}) \\
 &= P_{r_T}(H_0|H_1) \cdot P_{r_T}(H_1|H_2) \cdot P_{r_T}(H_2, H_3, \dots, H_{\tau^*}) \\
 &\quad \vdots \\
 &= P_{r_T}(H_{\tau^*}) \cdot \prod_{\tau=0}^{\tau^*-1} P_{r_T}(H_{\tau}|H_{\tau+1}) \tag{3.5}
 \end{aligned}$$

$$= \prod_{\tau=0}^{\tau^*-1} P_{r_T}(H_{\tau}|H_{\tau+1}). \tag{3.6}$$

Le passage de l'équation 3.5 à l'équation 3.6 est obtenu grâce à l'hypothèse que le MRCA est connu ; cette hypothèse implique $P_{r_T}(H_{\tau^*}) = 1$.

Probabilités de transition

Analysons maintenant la façon de calculer la probabilité conditionnelle $P_{r_T}(H_{\tau}|H_{\tau+1})$. Pour cela, rappelons les trois événements transitoires possibles, qui ont été décrits brièvement au chapitre précédent, afin de passer d'une génération à une autre dans un ARG et ce, du présent vers le passé. Il y a tout d'abord l'événement de coalescence (C) ; cet événement peut uniquement se produire entre deux séquences dont le matériel ancestral commun est identique. On distingue donc deux sortes d'événements de coalescence : la coalescence qui a lieu lorsque deux séquences identiques trouvent un ancêtre commun (voir un exemple sur la figure 3.2-(a)) et la coalescence qui a lieu lorsque deux séquences qui diffèrent uniquement par du matériel non-ancestral coalescent (voir un exemple sur la figure 3.2-(b)). Un événement de mutation (M), qui correspond en fait à changer un marqueur d'un statut muté ($\ll 1 \gg$) à non-muté ($\ll 0 \gg$), peut survenir uniquement s'il existe une seule séquence ayant un statut muté à ce marqueur. Rappelons que cette condition provient du fait que le modèle des sites infinis pour les mutations ne

permet pas que l'allèle à un locus donné mute plus d'une fois. De plus, les événements de mutation sur des marqueurs non-ancestraux ne sont pas pris en considération; un événement de mutation est illustré à la figure 3.2-(c). Finalement, un événement de recombinaison (R) peut se produire entre deux marqueurs d'une séquence uniquement s'il existe au moins un marqueur ancestral de chaque côté du point de recombinaison. Cette condition est causée par le fait qu'on ne considère pas les séquences constituées uniquement de matériel non-ancestral dans notre ARG, car ce type de séquences ne nous apporte aucune information sur la généalogie de notre échantillon. Un événement de recombinaison est illustré à la figure 3.2-(d). La notation suivante pour les événements de transition entre les états H_τ et $H_{\tau+1}$ sera utilisée pour la suite de ce mémoire :

- C_i : Coalescence entre deux séquences de type i ;
- C_{ij}^k : Coalescence des séquences i et j résultant en une séquence k ;
- $M_i^j(m)$: Mutation d'une séquence i au marqueur m résultant en une séquence j ;
- $R_i^{jk}(p)$: Recombinaison d'une séquence i dans l'intervalle p résultant en des séquences j et k .

Au chapitre précédent, nous avons montré que les probabilités que le prochain événement d'un ARG soit une coalescence, une mutation et une recombinaison, sachant qu'il a n séquences dans la génération considérée, sont respectivement :

$$\frac{n-1}{n-1+\theta+\rho}, \quad \frac{\theta}{n-1+\theta+\rho}, \quad \frac{\rho}{n-1+\theta+\rho}.$$

Ces probabilités doivent cependant être modifiées afin de prendre en considération la présence de matériel non-ancestral dans certains types de séquences; en effet, il a été mentionné au paragraphe précédent que la présence de tel matériel impose une restriction sur les événements de mutation et de recombinaison possibles entre deux générations consécutives d'un ARG.

Définissons donc quelques valeurs qui faciliteront le calcul de ces nouvelles probabilités. Posons n_i le nombre de séquences de type i présentes à l'état H_τ tel que $\sum_i n_i = n$, où n est le nombre total de séquences contenus dans H_τ . Posons $\alpha = (\sum_i n_i \cdot a_i)/nL$,

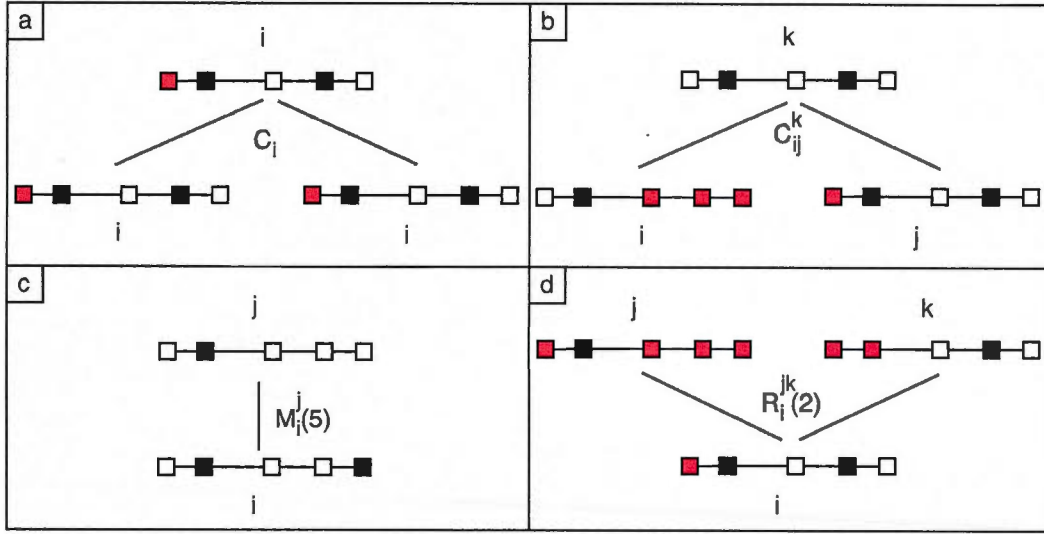


Figure 3.2 Exemple des différents événements de transition possibles, où les carrés roses représentent des marqueurs non-ancestraux. La figure (a) représente une coalescence entre deux séquences identiques, la figure (b) illustre une coalescence entre deux séquences différentes, un événement de mutation au marqueur 5 est illustré à la figure (c) et la figure (d) représente une recombinaison ayant lieu dans le deuxième intervalle d'une séquence de type i .

où a_i est le nombre de marqueurs ancestraux contenus dans une séquence de type i . La valeur α représente donc la proportion des marqueurs ancestraux parmi l'ensemble des marqueurs contenus dans un certain état H_τ ; le taux de mutation $n\theta/2$ devient donc $n\alpha\theta/2$. Nous allons poser $\beta = (\sum_i n_i \cdot r_i)/nr$, où r_i est la distance comprise entre les marqueurs ancestraux situés aux extrémités d'une séquence de type i et rappelons que r est la longueur totale d'une séquence. On peut donc interpréter β comme étant la proportion des portions de séquences où un événement de recombinaison est autorisé; le taux de recombinaison $n\rho/2$ devient donc $n\beta\rho/2$. Une fois adaptées à la présence de matériel non-ancestral, les trois probabilités précédentes deviennent :

$$P_\tau(C) = \frac{n-1}{n-1+\alpha\theta+\beta\rho}, \quad P_\tau(M) = \frac{\alpha\theta}{n-1+\alpha\theta+\beta\rho}, \quad P_\tau(R) = \frac{\beta\rho}{n-1+\alpha\theta+\beta\rho},$$

où l'indice τ indique que ces probabilités sont associées aux séquences présentes à l'état H_τ . La figure 3.3 illustre un exemple du calcul de certaines valeurs que nous venons de définir pour un échantillon contenant six séquences. Dans cet échantillon, les valeurs de α et β sont :

$$\alpha = \frac{\sum_{i=1}^4 n_i \cdot a_i}{nL} = \frac{(2 \cdot 2) + (1 \cdot 5) + (2 \cdot 4) + (1 \cdot 5)}{6 \cdot 5} = 0.73$$

$$\beta = \frac{\sum_{i=1}^4 n_i \cdot r_i}{nr} = \frac{(2 \cdot 1) + (1 \cdot 7) + (2 \cdot 6) + (1 \cdot 7)}{6 \cdot 7} = 0.67$$

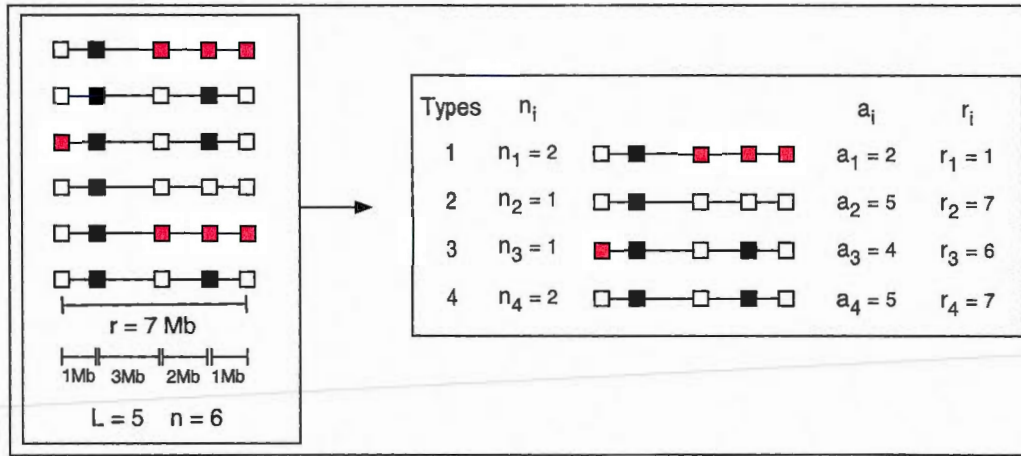


Figure 3.3 Illustration d'un échantillon contenant 6 séquences génétiques, chacune formée de 5 marqueurs et de longueur 7 Mb. Les valeurs n_i , a_i et r_i sont calculées pour chacun des 4 types de séquences présents dans l'échantillon.

Tous les outils sont maintenant en place afin de calculer la probabilité conditionnelle $P_{r\tau}(H_\tau | H_{\tau+1})$ pour chacun des événements de transition possibles. Il est important de souligner, afin de clarifier les calculs qui suivent, que les événements transitoires sont définis d'une génération à une autre du présent vers le passé, mais que la probabilité conditionnelle d'intérêt est calculée en regardant ces mêmes événements du passé vers le présent. Supposons tout d'abord que l'événement qui s'est produit entre les états H_τ et $H_{\tau+1}$ soit une coalescence. Considérons premièrement le cas où cette coalescence a eu lieu entre deux séquences de type i , nous allons représenter cet événement par : $H_{\tau+1} = H_\tau + C_i$. Un exemple d'un tel événement avec $i = 4$ est illustré à la figure

3.4(a). En supposant $|H_\tau| = n$, nous obtenons que $|H_{\tau+1}| = n - 1$ et que la séquence en moins dans $H_{\tau+1}$ est une séquence de type i . La probabilité conditionnelle recherchée est donc la probabilité qu'il y ait eu une coalescence entre les états H_τ et $H_{\tau+1}$, i.e $P_\tau(C)$, multipliée par la probabilité qu'une séquence de type i à l'état $H_{\tau+1}$ soit le résultat de cette coalescence, i.e $(n_i - 1)/(n - 1)$. Nous obtenons donc :

$$P_{rT}(H_\tau | H_{\tau+1} = H_\tau + C_i) = P_\tau(C) \cdot \frac{n_i - 1}{n - 1}.$$

Dans le cas où $H_{\tau+1} = H_\tau + C_{ij}^k$, nous obtenons :

$$P_{rT}(H_\tau | H_{\tau+1} = H_\tau + C_{ij}^k) = P_\tau(C) \cdot \frac{n_k + 1 + \delta_{ik} + \delta_{jk}}{n - 1}, \quad \text{où} \quad \delta_{ab} = \begin{cases} 1 & \text{si } a = b, \\ 0 & \text{sinon.} \end{cases}$$

Le numérateur $n_k + 1 + \delta_{ik} + \delta_{jk}$ représente le nombre de séquences de type k dans $H_{\tau+1}$ en prenant en considération que le type k résultant de la coalescence de i et j peut être différent de ces deux types ou le même que l'un d'entre eux. Un exemple de coalescence entre des séquences de type $i = 1$ et $j = 3$ résultant en une séquence de type $k = 4$ est illustré à la figure 3.4(b).

Si le prochain événement est une mutation d'une séquence de type i au marqueur m , résultant en une séquence de type j à l'état $H_{\tau+1}$, alors $H_{\tau+1} = H_\tau + M_i^j(m)$ et

$$P_{rT}(H_\tau | H_{\tau+1} = H_\tau + M_i^j(m)) = P_\tau(M) \cdot \frac{1}{\alpha L} \cdot \frac{n_j + 1}{n}.$$

La fraction $1/\alpha L$ représente la probabilité que la mutation soit survenue au marqueur m ; notons que la probabilité de mutation est supposée égale pour chaque marqueur. De plus, la valeur $n_j + 1$ correspond au nombre de séquences de type j dans $H_{\tau+1}$ et $n = |H_\tau| = |H_{\tau+1}|$, car le nombre de séquence total n'est pas modifié lors d'un événement de mutation. Un exemple d'un événement de mutation du marqueur $m = 5$ d'une séquence de type $i = 2$ résultant en une séquence de type $j = 5$ est présenté à la figure 3.4(c).

Finalement, considérons un événement de recombinaison dans l'intervalle p d'une séquence de type i , résultant en deux séquences de type j et k respectivement. Notons qu'un

exemple d'un tel événement avec $p = 2$, $i = 2$, $j = 1$ et $k = 5$ est illustré à la figure 3.4(d). On obtient alors $H_{\tau+1} = H_\tau + R_i^{jk}(p)$ et :

$$P_{r_T}(H_\tau | H_{\tau+1} = H_\tau + R_i^{jk}(p)) = P_\tau(R) \cdot \frac{r_p}{\beta r} \cdot \frac{(n_j + 1)(n_k + 1)}{(n + 1) \cdot n},$$

où r_p est la longueur de l'intervalle p et donc $r_p/\beta r$ représente la probabilité que le point de recombinaison ait été choisi dans l'intervalle p . De plus, il y a $n + 1$ séquences présentes à l'état $H_{\tau+1}$ et donc $(n + 1) \cdot n$ façons de choisir deux séquences aléatoirement pour les faire recombiner. Ainsi, la fraction $(n_j + 1)(n_k + 1)/(n + 1) \cdot n$ représente la probabilité d'avoir choisi aléatoirement un couple composé d'une séquence de type j et d'une séquence de type k .

Nous venons de voir la façon de calculer la probabilité d'un graphe $G = \{H_0, H_1, \dots, H_{\tau^*}\}$, la prochaine sous-section sera consacrée à la présentation de deux distributions proposées permettant de construire de tel graphe.

3.2.3 Échantillonnage pondéré : deux distributions proposées d'intérêt

Il a été mentionné à la section 3.2.1 que les ARGs employés dans le calcul de la vraisemblance du paramètre r_T sont générés à l'aide d'une distribution proposée $Q(G)$. De plus, il a été spécifié que la distribution utilisée doit avoir la particularité de générer des généalogies consistantes avec les données incluses dans $H_0|r_T$; l'astuce afin de construire des généalogies ayant cette particularité est de les construire à l'aide d'une chaîne de Markov dont l'état initial est $H_0|r_T$. De plus, dans notre contexte, les généalogies sont construites conditionnellement au paramètre r_T , on emploiera donc la notation $Q_{r_T}(G)$ afin de souligner cette condition. Notons que le choix d'une distribution proposée a un grand impact sur l'efficacité d'une méthode d'échantillonnage pondéré, il faut donc la choisir avec attention. Dans cette section, nous allons étudier les deux différentes distributions proposées que la méthode MapARG peut employer afin de construire les généalogies les plus probables possibles.

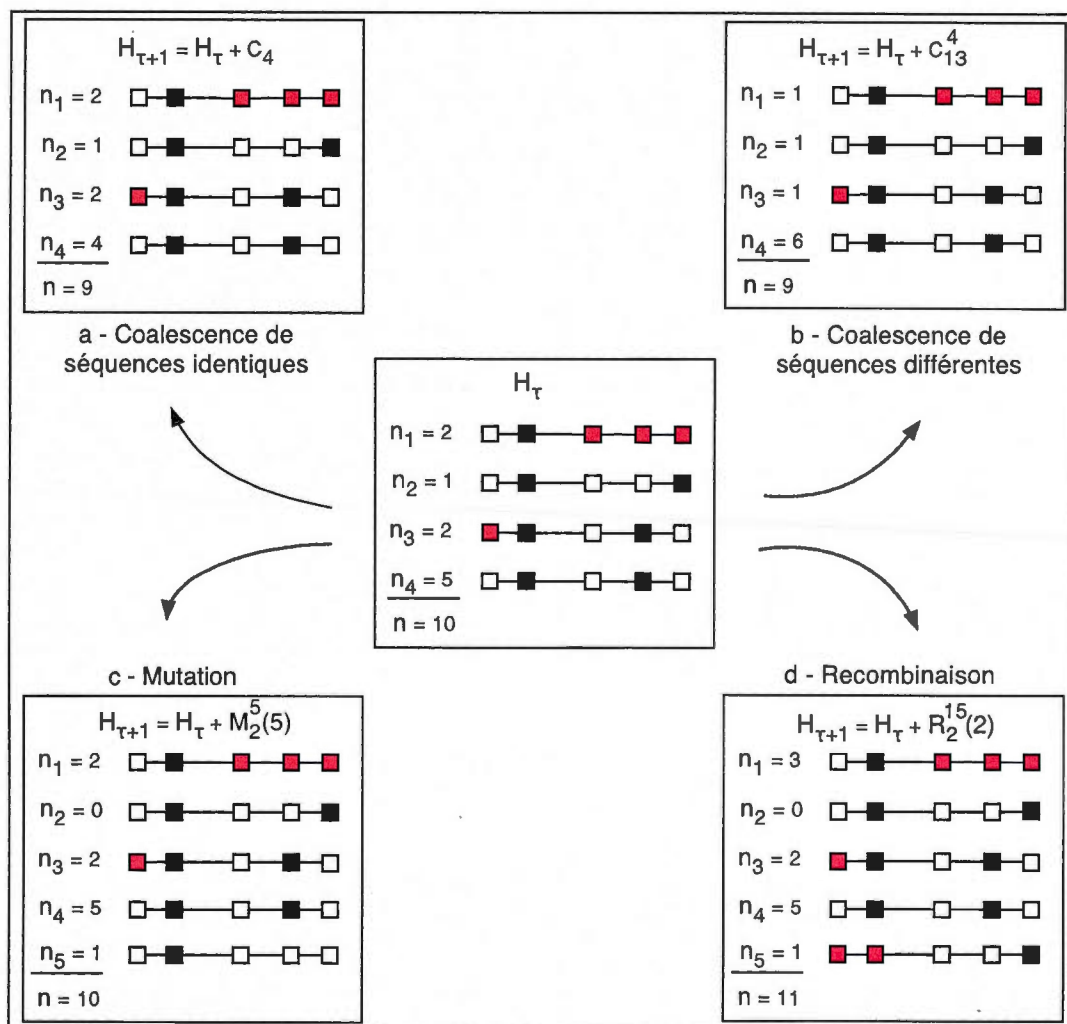


Figure 3.4 Illustration du passage de l'état H_τ à l'état $H_{\tau+1}$ pour un exemple de chaque événement possible. Un événement de coalescence entre deux séquences identiques est illustré en (a), nous retrouvons en (b) une coalescence entre deux séquences différentes, un événement de mutation est représenté en (c) et finalement un événement de recombinaison est illustré en (d).

Distribution proposée de Griffiths et Marjoram

La distribution proposée qui était originellement utilisée dans la méthode MapARG est une adaptation de celle proposée par Griffiths et Marjoram (1996). Cette distribution

$Q_{r_T}(G)$ construit un ARG à rebours dans le temps à l'aide d'une chaîne de Markov dont les probabilités de transition sont :

$$Q_{r_T}(H_{\tau+1}|H_{\tau}) = \frac{P_{r_T}(H_{\tau}|H_{\tau+1})}{\sum_{H_{\tau+1}} P_{r_T}(H_{\tau}|H_{\tau+1})},$$

où la somme présente au dénominateur est calculée sur tous les états $H_{\tau+1}$ tels que $P(H_{\tau}|H_{\tau+1}) \neq 0$, c'est-à-dire sur tous les états $H_{\tau+1}$ que l'on peut atteindre de l'état H_{τ} suite à un événement de coalescence, de mutation ou de recombinaison. En calculant $Q_{r_T}(G)$ d'une façon similaire à $P_{r_T}(G)$ (voir section 3.2.2), on obtient :

$$\begin{aligned} Q_{r_T}(G) &= Q_{r_T}(H_0, H_1, \dots, H_{\tau^*}) \\ &= Q_{r_T}(H_0) \cdot \prod_{\tau=0}^{\tau^*-1} Q_{r_T}(H_{\tau+1}|H_{\tau}) \\ &= Q_{r_T}(H_0) \cdot \prod_{\tau=0}^{\tau^*-1} \frac{P_{r_T}(H_{\tau}|H_{\tau+1})}{\sum_{H_{\tau+1}} P_{r_T}(H_{\tau}|H_{\tau+1})} \\ &= \prod_{\tau=0}^{\tau^*-1} \frac{P_{r_T}(H_{\tau}|H_{\tau+1})}{\sum_{H_{\tau+1}} P_{r_T}(H_{\tau}|H_{\tau+1})}. \end{aligned} \tag{3.7}$$

L'équation 3.7 est obtenue grâce au fait que la chaîne de Markov considérée débute à l'état H_0 ; le fait que cet état est fixe et connu nous permet de poser $Q_{r_T}(H_0) = 1$. En remplaçant respectivement $P_{r_T}(G)$ et $Q_{r_T}(G)$, dans l'équation 3.4 de la vraisemblance du paramètre r_T , par les équations 3.6 et 3.7 ; on obtient finalement la représentation suivante de la fonction de vraisemblance :

$$\begin{aligned} L(r_T) &\approx \frac{1}{M} \sum_{i=1}^M \left(\prod_{\tau=0}^{\tau^*-1} \frac{P_{r_T}(H_{\tau}^{(i)}|H_{\tau+1}^{(i)})}{P_{r_T}(H_{\tau}^{(i)}|H_{\tau+1}^{(i)}) / \sum_{H_{\tau+1}^{(i)}} P_{r_T}(H_{\tau}^{(i)}|H_{\tau+1}^{(i)})} \right) \\ &= \frac{1}{M} \sum_{i=1}^M \left(\prod_{\tau=0}^{\tau^*-1} \left(\sum_{H_{\tau+1}^{(i)}} P_{r_T}(H_{\tau}^{(i)}|H_{\tau+1}^{(i)}) \right) \right). \end{aligned}$$

Les résultats obtenus avec cette distribution sont satisfaisants, mais ils présentent une grande variabilité. Nous n'utiliserons donc pas cette distribution lors de l'utilisation de la méthode MapARG, mais plutôt celle présentée dans la sous-section suivante.

Distribution proposée de Fearnhead et Donnelly

Nous avons récemment implanté une adaptation de la distribution proposée de Fearnhead et Donnelly (2001) dans MapARG afin de pallier au problème de la variabilité mentionné ci-haut. Le changement par rapport à la distribution proposée précédente est au niveau des probabilités de transition de la chaîne de Markov. Fearnhead et Donnelly (2001) ont démontré que la distribution proposée optimale afin de construire des ARGs a pour probabilité de transition :

$$Q_{rT}(H_{\tau+1}|H_{\tau}) = P_{rT}(H_{\tau}|H_{\tau+1}) \cdot \frac{\pi(H_{\tau+1})}{\pi(H_{\tau})}, \quad (3.8)$$

où $\pi(H_{\tau})$ représente la probabilité qu'un échantillon contenant des séquences génétiques tirées au hasard parmi une population soit identique à l'échantillon contenu dans H_{τ} , lorsque l'on considère uniquement le matériel ancestral contenu dans H_{τ} . Il faut cependant approximer la probabilité $\pi(H_{\tau})$, celle-ci étant inconnue. Pour cela, nous devons tout d'abord définir $\pi(\cdot|H)$ comme étant la distribution conditionnelle du type de la dernière séquence tirée afin de former un échantillon sachant le type de toutes les autres séquences de cet échantillon. En posant H pour l'échantillon ne contenant que les séquences connues, on obtient alors :

$$\pi(\alpha|H) = \frac{\pi(\{H, \alpha\})}{\pi(H)},$$

où α représente un type de séquence quelconque. Il est possible d'écrire le quotient $\pi(H_{\tau+1})/\pi(H_{\tau})$ en fonction de cette distribution conditionnelle. En effet, supposons qu'un événement de coalescence entre deux séquences de type i soit survenu entre les états H_{τ} et $H_{\tau+1}$ et posons $H_{\tau+1} = H_{\tau} + C_i \equiv H_{\tau} - i$, c'est-à-dire que les séquences présentes à l'état $H_{\tau+1}$ sont exactement les mêmes séquences que celles à l'état H_{τ} à l'exception d'une séquence de type i en moins. On obtient alors l'équation suivante pour le quotient d'intérêt :

$$\frac{\pi(H_{\tau+1})}{\pi(H_{\tau})} = \frac{\pi(H_{\tau} - i)}{\pi(H_{\tau})} = \frac{\pi(H_{\tau} - i)}{\pi(H_{\tau} - i) \cdot \pi(i|H_{\tau} - i)} = \frac{1}{\pi(i|H_{\tau} - i)}.$$

En effet, l'expression $\pi(H_{\tau} - i)$ représente la probabilité de tirer au hasard parmi une population un échantillon de séquences génétiques identique à l'ensemble $H_{\tau} - i$

et l'expression $\pi(i|H_\tau - i)$ est équivalente à la probabilité que la prochaine séquence tirée d'une population pour compléter un échantillon connu et contenant les séquences formant l'ensemble $H_\tau - i$ soit une séquence de type i . Ainsi, on obtient bien que $\pi(H_\tau - i) \cdot \pi(i|H_\tau - i) = \pi(H_\tau)$.

Avec une méthode similaire, on obtient les expressions suivantes pour les trois autres événements de transition :

$$\frac{\pi(H_{\tau+1})}{\pi(H_\tau)} = \begin{cases} \frac{\pi(k|H_\tau - i - j)}{\pi(i|H_\tau - i)\pi(j|H_\tau - i - j)} & \text{si } H_{\tau+1} = H_\tau + C_{ij}^k, \\ \frac{\pi(j|H_\tau - i)}{\pi(i|H_\tau - i)} & \text{si } H_{\tau+1} = H_\tau + M_i^j, \\ \frac{\pi(j|H_\tau - i)\pi(k|H_\tau - i + j)}{\pi(i|H_\tau - i)} & \text{si } H_{\tau+1} = H_\tau + R_i^{jk}. \end{cases}$$

Encore une fois, la distribution $\pi(\cdot|H)$ est inconnue, il existe toutefois une approximation efficace ($\hat{\pi}(\cdot|H)$) qui possède les deux caractéristiques suivantes :

- les types de séquences ayant une grande fréquence d'apparition dans l'échantillon H ont une probabilité considérable dans la distribution $\hat{\pi}(\cdot|H)$;
- considérons deux types de séquences α et β qui n'apparaissent pas dans H et supposons qu'une séquence de type α possède du matériel génétique plus similaire à celui des séquences présentes dans H qu'une séquence de type β , nous avons alors l'inégalité $\hat{\pi}(\alpha|H) > \hat{\pi}(\beta|H)$.

La façon d'approximer efficacement cette distribution conditionnelle est présentée à l'annexe A (page 115).

Avec cette distribution proposée, la fonction de vraisemblance de l'équation 3.4 peut s'écrire comme

$$\begin{aligned} L(r_T) &\approx \frac{1}{M} \sum_{i=1}^M \left(\prod_{\tau=0}^{\tau^*-1} \frac{P_{r_T}(H_\tau^{(i)}|H_{\tau+1}^{(i)})}{Q_{r_T}(H_{\tau+1}^{(i)}|H_\tau^{(i)})} \right) \\ &= \frac{1}{M} \sum_{i=1}^M \left(\prod_{\tau=0}^{\tau^*-1} \frac{P_{r_T}(H_\tau^{(i)}|H_{\tau+1}^{(i)})}{P_{r_T}(H_\tau^{(i)}|H_{\tau+1}^{(i)}) \cdot \pi(H_{\tau+1}^{(i)})/\pi(H_\tau^{(i)})} \right) \\ &= \frac{1}{M} \sum_{i=1}^M \left(\prod_{\tau=0}^{\tau^*-1} \frac{\pi(H_\tau^{(i)})}{\pi(H_{\tau+1}^{(i)})} \right). \end{aligned}$$

Les résultats obtenus avec cette distribution proposée sont comparables à ceux obtenus

avec la distribution de Griffiths et Marjoram, mais ils ont l'avantage d'être beaucoup moins variables. Nous emploierons donc cette distribution lors de l'utilisation de la méthode MapARG.

3.2.4 Vraisemblance composite et conditionnelle

Un problème non-négligeable de la méthode décrite dans les sections précédentes est l'énorme temps de calcul informatique nécessaire à son application. En effet, la simulation de graphes de recombinaison ancestraux peut être très longue, surtout pour de longues séquences, car plus la longueur des séquences augmente, plus les événements de recombinaison sont fréquents dans la généalogie, ce qui a pour conséquence d'augmenter le nombre de générations nécessaires avant l'atteinte d'un ancêtre commun. Une solution à ce problème est d'utiliser des vraisemblances composites et conditionnelles lors de l'estimation du paramètre r_T ; ce type de vraisemblance sera calculé à l'aide de généalogies construites à partir d'un sous-ensemble de marqueurs, le temps de simulation d'un graphe sera donc diminué. L'objectif de cette section est de présenter brièvement la fonction de vraisemblance composite et conditionnelle utilisée afin d'estimer le paramètre r_T .

L'idée est de calculer la vraisemblance du paramètre r_T à l'aide d'ARGs construits à partir d'un sous-ensemble de marqueurs, plutôt qu'avec des ARGs construits à partir de séquences complètes contenues dans l'échantillon. Un sous-ensemble de marqueurs consécutifs et connus sur une séquence est appelé une *fenêtre de marqueurs* et celle-ci contient un nombre fixe de $d - 1$ marqueurs. La figure 3.5(a) illustre la façon de diviser une séquence contenant $L - 1 = 6$ marqueurs connus en fenêtre de $d - 1 = 4$ marqueurs. En construisant des généalogies à partir des marqueurs contenus dans une fenêtre, il est possible d'évaluer la vraisemblance marginale de r_T à chacune des valeurs candidates pour lesquelles le TIM est inféré dans cette fenêtre, et ce, avec la même méthode que précédemment, à l'exception que les séquences ne sont plus formées de L marqueurs, mais bien de d marqueurs. Il est important de remarquer que le TIM inféré à une distance π_i du premier marqueur d'une séquence peut appartenir à plusieurs

fenêtres; par exemple, tel qu'illustré à la figure 3.5(b), le TIM inféré à une distance π_2 du premier marqueur d'une séquence, c'est-à-dire au milieu du deuxième intervalle de la séquence, est compris dans les deux premières fenêtres. La vraisemblance marginale de r_T pour une fenêtre donnée dépend donc des marqueurs contenus dans cette fenêtre. Ainsi, cette vraisemblance marginale est en fait une vraisemblance conjointe du paramètre r_T et des marqueurs contenus dans la fenêtre. La fonction d'intérêt étant la vraisemblance du paramètre r_T , nous allons utiliser une vraisemblance marginale conditionnelle aux marqueurs contenus dans la fenêtre considérée; la méthode classique permettant de combiner les différentes vraisemblances marginales conditionnelles évaluées à une même valeur candidate est de calculer une moyenne géométrique avec celles-ci; le développement formel de cette méthode est présenté dans ce qui suit.

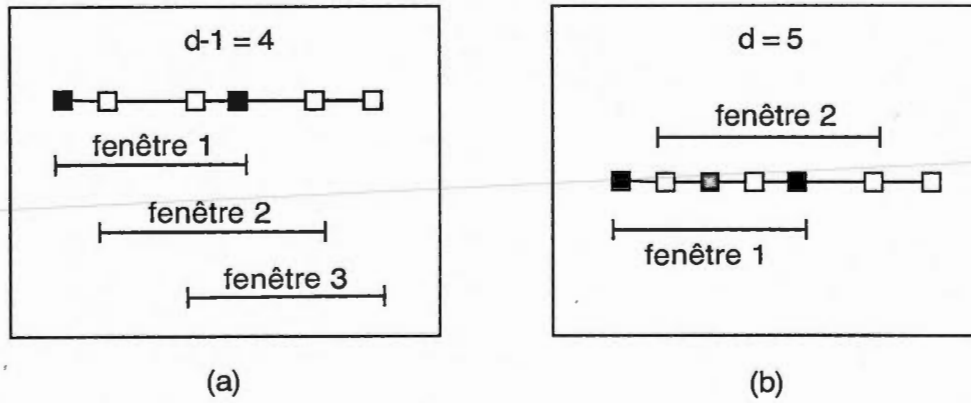


Figure 3.5 (a) Illustration de la division d'une séquence contenant $L-1 = 6$ marqueurs connus en $D = 3$ fenêtres de $d-1 = 4$ marqueurs chacune. (b) Illustration de la séquence présentée en (a) à laquelle nous avons inféré un marqueur correspondant au TIM au milieu du deuxième intervalle; ce marqueur est inclus dans les fenêtres 1 et 2.

Afin de formaliser la fonction de vraisemblance composite conditionnelle à calculer, introduisons quelques notations. Posons D le nombre de fenêtres obtenues en divisant les $L-1$ marqueurs d'une séquence en fenêtres contenant $d-1$ marqueurs. Pour $j = 1, \dots, D$, dénotons H_{0j} l'ensemble des séquences formées des marqueurs inclus dans la fenêtre j et du marqueur correspondant au TIM; les séquences incluses dans H_{0j}

contiennent donc d marqueurs. Dénotons de plus H_{0j}^* l'ensemble des séquences formées uniquement des marqueurs inclus dans la fenêtre j ; les séquences incluses dans H_{0j}^* contiennent donc $d - 1$ marqueurs. Nous allons de plus poser p pour l'intervalle formé par les p^e et $(p + 1)^e$ marqueurs connus d'une séquence; p prend donc des valeurs entre 1 et $L - 2$. Nous obtenons ainsi que l'intervalle p est inclus dans une fenêtre j uniquement si $\underline{J}(p) \leq j \leq \bar{J}(p)$, où $\underline{J}(p) = \max(1, p - d + 3)$ et $\bar{J}(p) = \max(p, L - d + 1)$.

Nous définissons la vraisemblance marginale conditionnelle aux marqueurs contenus dans une fenêtre j , évaluée à une valeur candidate π_i et pour un intervalle p quelconque de la façon suivante :

$$L_{j,p}(r_T = \pi_i | H_{0j}^*) = \begin{cases} \frac{P(H_{0j} | r_T)}{P(H_{0j}^*)} = \frac{L_j(r_T)}{P(H_{0j}^*)} & \begin{array}{l} \text{si le TIM inféré à une distance} \\ \pi_i \text{ du premier marqueur ap-} \\ \text{partient à l'intervalle } p, \end{array} \\ 1 & \text{sinon.} \end{cases}$$

La fonction de vraisemblance composite conditionnelle du paramètre r_T évaluée à une valeur candidate π_i peut donc s'écrire comme

$$CCL(r_T = \pi_i) = \prod_{p=1}^{L-2} \left(\prod_{j=\underline{J}(p)}^{\bar{J}(p)} L_{j,p}(r_T = \pi_i | H_{0j}^*) \right)^{w_p},$$

où w_p représente l'inverse du nombre de fenêtres contenant l'intervalle p , c'est-à-dire $w_p = 1/(\bar{J}(p) - \underline{J}(p) + 1)$. Notons que cette approche a été proposée dans Larribe et Lessard (2008).

3.2.5 Inférence d'allèle au marqueur correspondant au TIM

Dans la présentation du modèle à la section 3.2.1, nous avons vu que chaque séquence de notre échantillon possède un marqueur, correspondant au TIM, dont l'allèle est inconnu. Cette section a pour but d'expliquer la façon d'inférer le génotype («0» ou «1») à ce marqueur en utilisant le phénotype (cas ou témoin) qui lui est associé. Précisons

tout d'abord que la présence d'un allèle non-muté («0») à ce marqueur indique que la séquence provient d'un individu non-porteur de la mutation, tandis que la présence d'un allèle muté («1») signifie que cette séquence provient d'un individu porteur de la mutation. Il est important de remarquer que l'inférence d'un allèle au marqueur correspondant au TIM a un rôle crucial dans l'estimation de la position de celui-ci. En effet, les probabilités de transition d'un graphe sont non seulement conditionnelles à la position à laquelle le marqueur correspondant au TIM a été ajouté, mais aussi aux allèles présents à ce marqueur. Ainsi une mauvaise inférence impliquera une mauvaise évaluation de la vraisemblance $L(r_T)$ et donc possiblement le choix d'un estimateur biaisé.

Introduisons maintenant quelques termes et notations afin de faciliter la compréhension de cette section. Posons F la fonction de pénétrance nous indiquant la probabilité d'observer un certain phénotype en fonction du génotype qui lui est associé. Nous allons supposer cette fonction connue pour notre maladie d'intérêt et la noter $F = (f_0, f_1)$, où $0 \leq f_i \leq 1$ représente la probabilité qu'une séquence porteuse de i copie de la mutation provienne d'un individu atteint par la maladie ($i = 0, 1$). Lorsque $f_0 \neq 0$, on dit qu'il y a présence de phénocopie, c'est-à-dire qu'il y a des séquences non-porteuses de la mutation qui proviennent d'individu atteint par la maladie ; celle-ci n'est donc pas uniquement influencée par la mutation et elle peut donc être causée par d'autres facteurs, tels que des facteurs environnementaux. Dans le cas où $f_1 \neq 1$, on dit qu'il y a pénétrance incomplète ; certaines séquences possédant la mutation ne proviennent pas d'individus atteints par la maladie. La grande majorité des maladies que l'on retrouve de nos jours présente des pénétrances incomplète et/ou des phénocopies, il est donc très intéressant de développer des méthodes de cartographie génétique pour de telles maladies.

Lorsque l'on considère une maladie ayant une fonction de pénétrance $F = (0, 1)$, il est très facile d'inférer le génotype à partir du phénotype. En effet, dans cette situation, toutes les séquences non-porteuses de la mutation, c'est-à-dire possédant l'allèle «0», proviennent d'individus qui sont des témoins, tandis que toutes les séquences porteuses de la mutation (allèle «1») proviennent d'individus qui sont des cas. La méthode Map-

ARG, dont la performance est en partie liée à cette inférence, est donc très efficace pour de telles maladies. Cependant, la présence de phénocopie et/ou de pénétrance incomplète complique grandement l'inférence du génotype à partir du phénotype; un algorithme EM (Boucher, 2009) doit alors être utilisé afin d'estimer le génotype des séquences à l'aide de leur phénotype. Malgré l'utilisation d'un tel algorithme, l'inférence demeure difficile et parfois inefficace; la méthode est donc moins performante pour les maladies ayant des fonctions de pénétrance complexes.

3.2.6 Exemple de résultat

La figure 3.6 illustre un exemple de résultat obtenu avec la méthode MapARG pour des données simulées dont la description sera faite au début du chapitre 5. Notons que la distribution proposée qui a été utilisée pour obtenir ce résultat est celle de Fearnhead et Donnelly (2001).

3.3 Méthode Margarita

Cette section a pour but de présenter brièvement la méthode de cartographie génétique fine Margarita (Minichiello and Durbin, 2006).

3.3.1 Le modèle

Chaque séquence d'un échantillon est formée de L marqueurs connus, c'est-à-dire que nous connaissons la position et l'allèle de chacun de ces marqueurs. De plus, on suppose que l'emplacement du marqueur sur lequel le TIM est apparu est tout près de l'un de ces L marqueurs. La méthode Margarita permet d'estimer près de quel marqueur, parmi les L présents sur les séquences d'un échantillon, le TIM est apparu. L'idée générale afin d'obtenir une telle estimation est de tester pour chaque marqueur $m \in \{1, 2, \dots, L\}$ d'une séquence, l'association entre les allèles inférés à un marqueur m' situé tout près de m et le phénotype des séquences de notre échantillon. Notons que l'association est testée à l'aide d'un test du χ^2 et que la méthode utilisée afin d'inférer des allèles à un

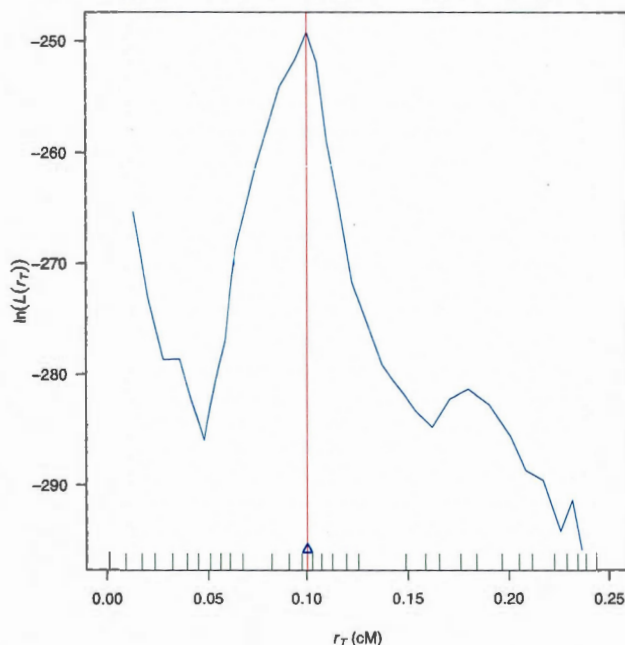


Figure 3.6 Graphique représentant le logarithme de la fonction de vraisemblance (ligne bleue). La ligne verticale rouge indique la vraie position de la mutation sur la séquence génétique et le triangle bleu l'estimation de cette position faite par MapARG. Nous observons donc que l'estimation faite par MapARG pour cet exemple est très bonne puisque le triangle bleu est superposé à la ligne rouge.

marqueur m' sera expliquée à la section 3.3.3. Une fois les tests d'association effectués, il est possible à l'aide de tests de permutation d'évaluer une valeur-p pour chacun des marqueurs dont l'association avec le phénotype a été testée ; l'estimation de la position du TIM utilisée par la méthode Margarita est la position du marqueur ayant la plus petite valeur-p.

3.3.2 Graphe de recombinaison ancestral et arbre partiel

Les graphes de recombinaison ancestraux utilisés par la méthode Margarita ne sont pas simulés à l'aide du processus de coalescence avec recombinaison, comme c'est le cas avec la méthode MapARG, mais plutôt à l'aide d'un algorithme heuristique. De plus, tous les

graphes simulés sont considérés équiprobables, la façon de calculer la probabilité d'un ARG n'est donc plus une question d'intérêt avec cette approche. Cette section est donc consacrée à la présentation de cet algorithme heuristique et à l'introduction de la notion d'arbres partiels.

L'algorithme heuristique développé par Minichiello et Durbin (2006) permet la construction de graphes du présent vers le passé, et ce, encore une fois, afin de s'assurer d'obtenir des généalogies consistantes avec les données. De plus, les événements de transition de coalescence et de mutation possibles entre deux générations consécutives sont les mêmes que dans la méthode MapARG. En effet, les coalescences ne sont permises qu'entre deux séquences dont le matériel ancestral commun est identique, et un événement de mutation peut survenir à un marqueur m uniquement s'il existe une unique séquence possédant un allèle muté à ce marqueur. Cette dernière condition résulte du fait que le modèle de mutation utilisé est celui des sites infinis. La particularité de cet algorithme heuristique est au niveau des événements de recombinaison ; ils ne sont pas définis de la même façon que dans le processus de coalescence avec recombinaison. En effet, dans cet algorithme, les événements de recombinaison sont définis de façon à simplifier la généalogie simulée, c'est-à-dire de façon à ce que le temps avant l'atteinte du MRCA soit le plus court possible. Notons que cette façon de faire est logique, puisque les généalogies dont le temps avant l'atteinte du MRCA est relativement court ont tendance à être probables. Mentionnons de plus qu'un événement de recombinaison peut survenir dans une généalogie générée avec cet algorithme heuristique uniquement lorsqu'aucun événement de coalescence et de mutation n'est possible.

Nous allons maintenant expliquer en détail la façon de simuler un événement de recombinaison, que nous noterons $R_i^{jk}(p)$, dans une généalogie afin de passer d'un état H_τ à un état $H_{\tau+1}$. Notons que nous utilisons la même notation que celle employée afin de décrire la méthode MapARG à la section 3.2, ainsi $R_i^{jk}(p)$ représente la recombinaison d'une séquence de type i dans l'intervalle p résultant en des séquences de type j et k ; nous allons donc décrire la façon de choisir le type i de la séquence recombinante et l'intervalle p de recombinaison. Mentionnons premièrement, qu'afin de simplifier la généalogie

simulée, l'événement $R_i^{jk}(p)$ devra avoir pour effet de permettre un événement de coalescence à la génération $\tau + 1$ si c'est possible, sinon à la génération $\tau + 2$. Afin de simuler un tel événement, il faut tout d'abord choisir deux séquences, dont nous noterons les types par q et r , présentes dans l'ensemble H_τ et possédant les mêmes allèles pour un sous-ensemble de marqueurs consécutifs. Ensuite, nous posons (a, b) l'intervalle comprenant ce sous-ensemble de marqueurs : a et b représentant respectivement le premier et dernier marqueur de celui-ci. Il y a alors deux points de recombinaison possibles : un dans l'intervalle $a - 1$ si $a \neq 1$ et l'autre dans l'intervalle b si $b \neq L$, où L représente le nombre total de marqueurs sur une séquence. Dans la situation où il y a uniquement un de ces deux points de recombinaison valide (voir figure 3.7 Cas (a)), supposons que c'est celui dans l'intervalle $a - 1$ (et donc que $b = L$) ; il suffit alors de choisir une séquence recombinante de type q ou r et de la faire recombiner dans l'intervalle $a - 1$. Supposons que l'événement obtenu est $R_q^{ef}(a - 1)$, on obtient alors que la séquence de type f pourra coalescer avec la séquence de type r à la génération $\tau + 1$. Dans la situation où les deux points de recombinaison sont valides (voir figure 3.7 Cas (b)), c'est-à-dire que $a \neq 1$ et $b \neq L$, un événement de coalescence entre les séquences de type f et r ne sera pas possible et il faudra que le prochain événement de la généalogie soit encore une recombinaison. Cette recombinaison doit avoir lieu sur une séquence de type f dans l'intervalle b , posons cet événement $R_f^{tz}(b)$, on obtient finalement que les séquences de types r et t pourront coalescer à la génération $\tau + 2$.

L'innovation de l'algorithme heuristique employé dans la méthode Margarita est dans le choix de l'événement de transition ; en effet, celui-ci n'est plus fait proportionnellement à la probabilité de chacun des événements possibles, c'est-à-dire en respectant la théorie de la coalescence avec recombinaison, mais d'une façon heuristique. Voici les règles heuristiques à suivre afin de construire un ARG avec cet algorithme :

1. un événement de recombinaison peut se produire uniquement si aucun événement de coalescence et de mutation n'est possible ;
2. lorsque plusieurs événements de coalescence et de mutation sont possibles, le prochain événement à survenir dans l'ARG est choisi aléatoirement parmi ceux-ci ;

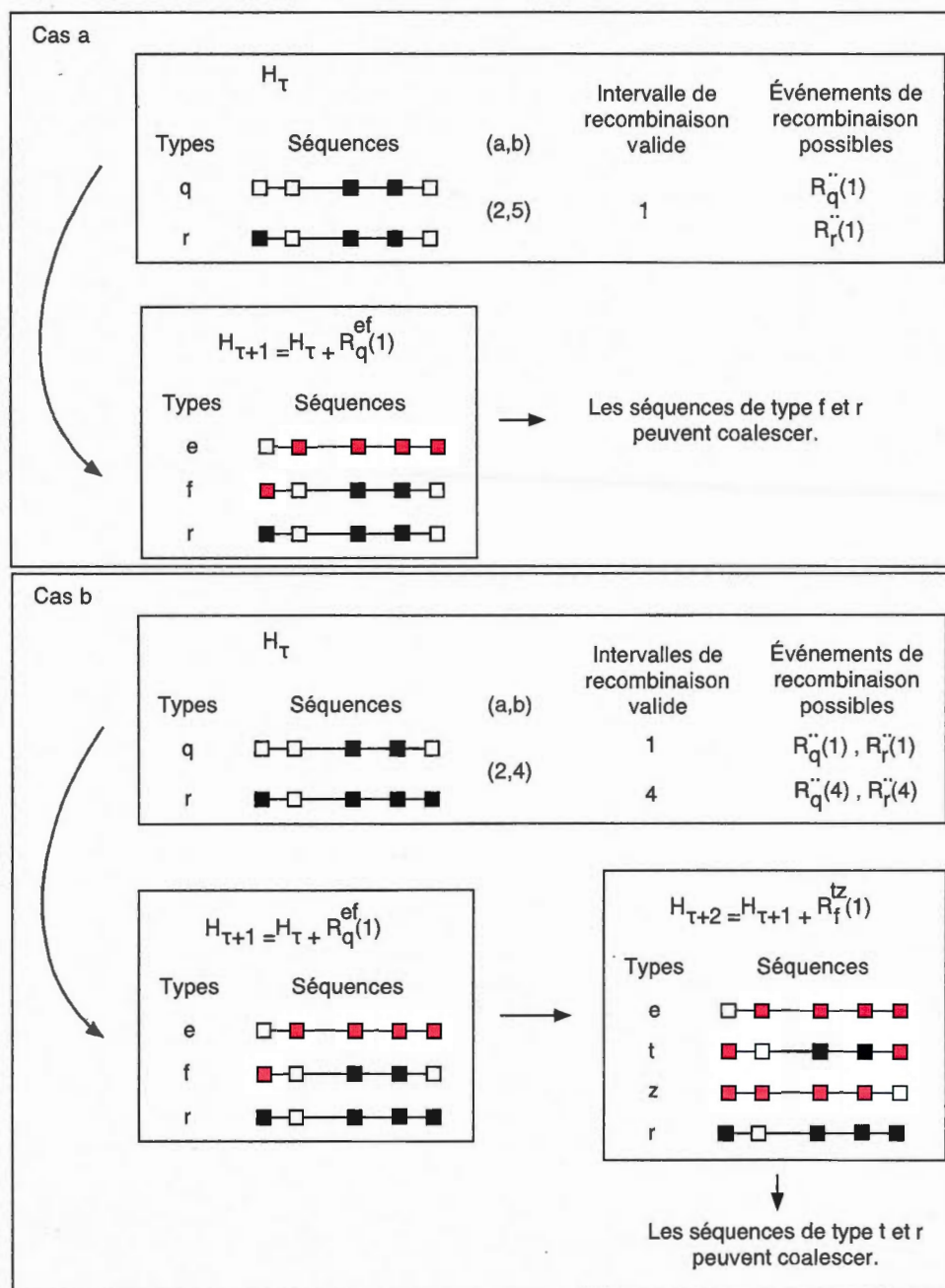


Figure 3.7 Illustration de la simulation d'un événement de recombinaison avec l'algorithme heuristique. Le cas (a) représente la situation où il y a un unique intervalle de recombinaison valide et le cas (b) représente celle où les deux intervalles de recombinaison sont valides.

3. un événement de coalescence n'est pas permis entre deux séquences dont tout le matériel commun est non-ancestral ;
4. lorsque l'événement transitoire doit être une recombinaison, les deux séquences permettant de créer cet événement sont choisies telles que l'intervalle (a, b) , contenant un sous-ensemble de marqueurs consécutifs pour lesquels les allèles de ces deux séquences sont les mêmes, soit de longueur maximale avec une probabilité de 0.9 et de façon aléatoire avec une probabilité de 0.1 ;
5. le premier événement de coalescence suivant un événement de recombinaison R_i^{jk} doit faire intervenir une séquence de type j ou k .

Introduisons maintenant la notion d'arbre partiel, qui sera nécessaire à la compréhension de la méthode Margarita et de la méthode DMap présentée dans le prochain chapitre. Un graphe de recombinaison ancestral pour un échantillon de séquences contenant L marqueurs contient L arbres partiels (appelés aussi arbres marginaux) ; un arbre partiel est simplement la généalogie d'un unique marqueur m ($m \in \{1, 2, \dots, L\}$). Il est très facile d'extraire d'un ARG l'arbre partiel d'un marqueur m ; en effet, il suffit de conserver uniquement les lignées de l'ARG pour lesquelles l'allèle à ce marqueur est ancestral. Notons que les différences entre les arbres partiels extraits d'un même ARG sont le résultat des événements de recombinaison ; en effet, une recombinaison crée une bifurcation dans un graphe telle que les arbres partiels des marqueurs à gauche du point de recombinaison contiendront la branche pointant vers la gauche et ceux des marqueurs à sa droite contiendront la branche pointant vers la droite. La figure 3.8 illustre un ARG construit à l'aide de l'algorithme heuristique pour un échantillon de quatre séquences de longueur $L = 4$ ainsi que les arbres partiels pour le deuxième et le quatrième marqueur.

3.3.3 Inférence du TIM et algorithme utilisé par Margarita

Il a été mentionné à la section 4.1 que la méthode Margarita permet d'estimer près de quel marqueur parmi les L connus est situé le TIM. Pour cela, il faut tester pour chaque marqueur $m \in \{1, \dots, L\}$ l'association entre les allèles d'un marqueur m' situé

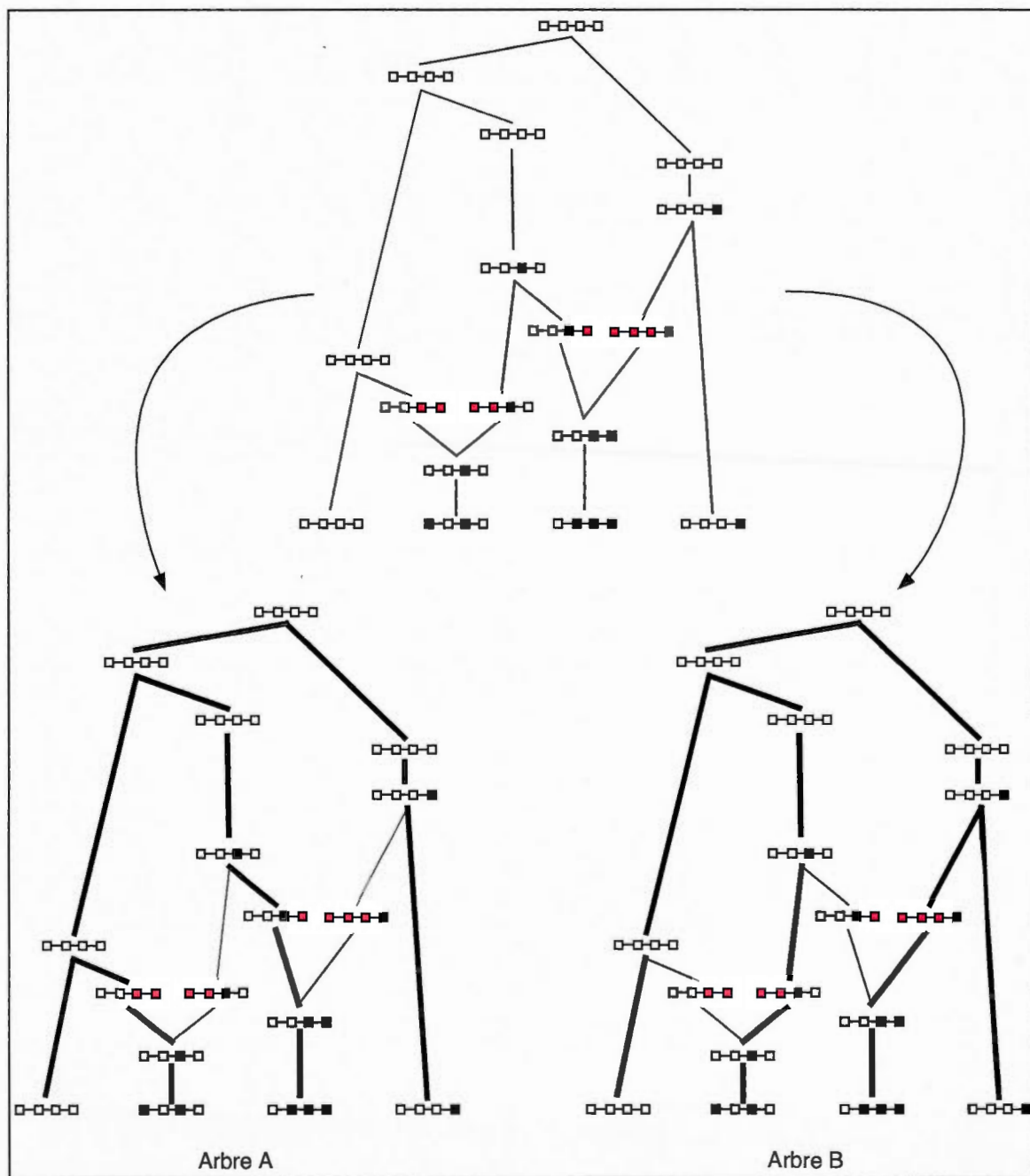


Figure 3.8 Illustration d'un graphe de recombinaison ancestral construit avec l'algorithme heuristique (en haut), les branches en gras de l'arbre A forment l'arbre partiel du deuxième marqueur et celles de l'arbre B forment l'arbre partiel du quatrième marqueur.

très près de m et le phénotype des séquences de l'échantillon, et ce, à l'aide d'un test du χ^2 . Pour cela, il est d'abord nécessaire d'inférer des allèles au marqueur m' à évaluer ; cette inférence se fait à l'aide de l'arbre partiel du marqueur m . En effet, en choisissant m' très près de m , on obtient que la probabilité qu'un événement de recombinaison soit survenu entre ces deux marqueurs est quasiment nulle, on peut donc supposer que les arbres partiels des marqueurs m et m' sont les mêmes. L'inférence d'un allèle primitif ou mutant au marqueur m' de chaque séquence d'un échantillon se fait en ajoutant la mutation causale sur une des branches de l'arbre partiel du marqueur m' et en observant les séquences de l'échantillon qui en héritent. Nous obtenons ainsi une division des séquences en deux catégories : porteuses ou non-porteuses du TIM. La figure 3.9 illustre ce phénomène et le tableau 3.1 représente la table de contingence obtenue suite à l'inférence. Notons que la nouvelle méthode qui sera présentée au chapitre suivant utilise une méthode d'inférence très similaire à celle décrite dans ce paragraphe.

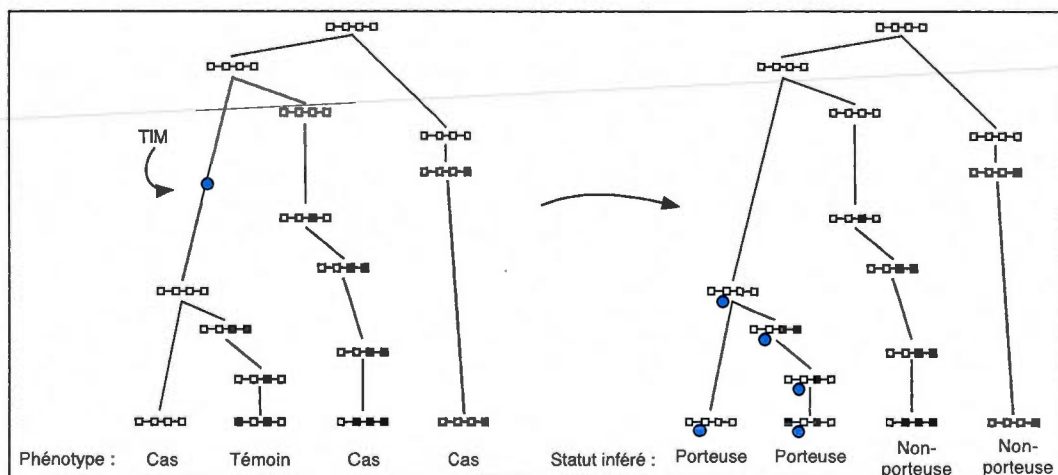


Figure 3.9 À gauche, nous retrouvons l'illustration d'un arbre partiel pour le deuxième marqueur ($m = 2$) à l'aide duquel on infère à chacune des séquences de notre échantillon un statut de porteuse ou non-porteuse de la mutation au marqueur m' . Nous avons ajouté une mutation (cercle bleu) sur une de ces branches et l'arbre de droite représente la façon dont la mutation s'est propagée dans l'arbre.

Statut / Phénotype	Cas	Témoin	Totaux
Porteuse	1	1	2
Non-porteuse	2	0	2
Totaux	3	1	4

Tableau 3.1 Table de contingence obtenue suite à l'inférence illustrée à la figure 3.9

L'algorithme utilisé par la méthode Margarita afin de calculer un score d'association entre les allèles des marqueurs à tester et le phénotype des séquences est :

1. Générer M graphes de recombinaison ancestraux à l'aide de l'algorithme heuristique présenté à la section 3.3.2.
2. Pour $g = 1, \dots, M$ faire :
 - Extraire les L arbres partiels contenu dans l'ARG g .
 - Pour $m_g = 1, \dots, L$ faire :
 - (a) Pour chacune des branches $b = 1, \dots, B_{m_g}$ de l'arbre partiel du marqueur m_g , faire :
 - Poser le TIM sur la branche b ; ceci a pour effet d'inférer un statut de porteuse ou de non-porteuse du TIM à chaque séquence de l'échantillon.
 - Tester à l'aide d'un test du χ^2 l'association entre les statuts inférés et le phénotype (cas ou témoin) des séquences. Poser S_b la statistique du χ^2 obtenue.
 - (b) Le score de l'arbre partiel du marqueur m_g est $S_{m_g} = \max\{S_b | b \in \{1, \dots, B_{m_g}\}\}$.
3. Le score d'association pour chaque marqueur m est $S_m = (\sum_{g=1}^M S_{m_g})/M$.

Une fois les scores d'association obtenus grâce à cet algorithme, il est possible de tester s'ils sont significatifs à l'aide de l'obtention d'une valeur-p pour chaque marqueur. Ces valeurs-p peuvent être obtenues aisément à l'aide d'un test de permutation consistant à appliquer plusieurs fois l'algorithme précédent à des données où l'assignement des cas et des témoins a été permuté. L'estimation de la position du TIM est la position du marqueur ayant obtenu la plus petite valeur-p.

3.3.4 Exemple de résultat

La figure 3.10 illustre un exemple de résultat obtenu avec la méthode Margarita pour des données simulées (Minichiello and Durbin, 2006).

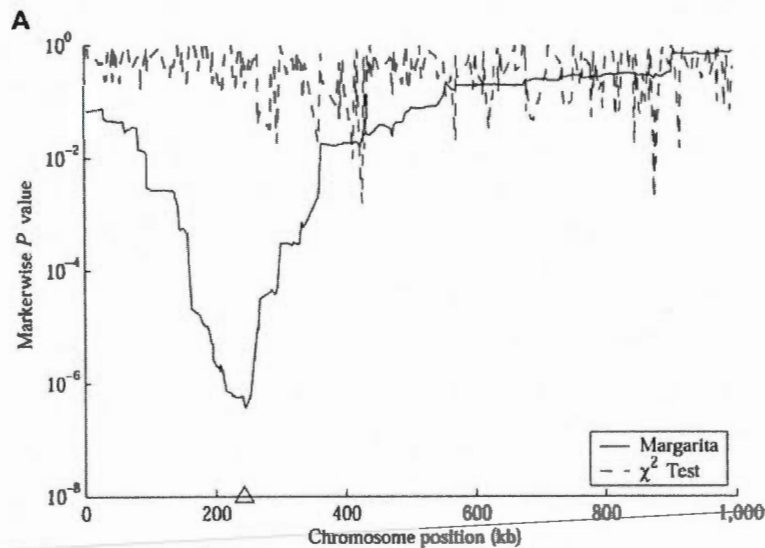


Figure 3.10 Graphique représentant les valeur-p obtenues pour chaque test d'association entre une position le long de la séquence génétique et la mutation causale. La ligne pleine représente les résultats obtenus avec la méthode Margarita, tandis que la ligne brisée représente ceux obtenue avec un test standard du χ^2 . Le triangle indique la vraie position de la mutation causale sur la séquence génétique; nous pouvons donc observer que le marqueur ayant la plus petite valeur-p selon la méthode Margarita est situé très près de la mutation causale et que l'estimation faite par cette méthode est donc très bonne pour cet exemple (tiré de Minichiello et Durbin, 2006).

3.4 Méthode LATAG

La méthode de cartographie génétique LATAG (Zöllner and Pritchard, 2005), est une méthode à deux niveaux. En effet, en plus de permettre l'estimation de la position sur

une séquence génétique d'une mutation influençant une maladie, elle peut être utilisée afin de déterminer s'il est probable qu'une courte région d'un chromosome contienne une telle mutation. Puisque nous nous intéressons principalement à la cartographie génétique fine, cette section présentera uniquement la méthodologie afin d'estimer la position d'un TIM. De plus, il est important de mentionner que la méthode peut être utilisée dans des cas où le phénotype associé à la maladie d'intérêt est quantitatif plutôt que qualitatif; ceci signifie que les phénotypes possibles pourraient être, par exemple, des taux de cholestérol plutôt que d'être uniquement les catégories cas et témoin. Nous allons cependant présenter uniquement la méthode pour des maladies dont le phénotype est qualitatif.

3.4.1 Le modèle

Le modèle sous-jacent à la méthode LATAG considère des échantillons contenant n séquences génétiques formées de L marqueurs connus, c'est-à-dire que nous connaissons la position et l'allèle de chacun de ces marqueurs. Une approche bayésienne est employée dans ce modèle afin d'estimer la position, notée x , d'une mutation causale; pour cela, la probabilité a posteriori de chaque élément d'un ensemble de positions possibles pour le TIM, noté $X = \{x_1, x_2, \dots, x_k\}$, est calculée et l'estimateur utilisé est la position $x_j \in X$ ayant la plus grande probabilité a posteriori. En posant H_0 l'ensemble contenant les n séquences formées de L marqueurs de notre échantillon et $\Phi = (\phi_1, \phi_2, \dots, \phi_n)$ le vecteur contenant le phénotype de chacune de ces séquences, on obtient la distribution a posteriori suivante (Zöllner and Pritchard, 2005) :

$$P(x_j | \Phi, H_0) = \frac{P(\Phi, H_0 | x_j) P(x_j)}{\int_X P(\Phi, H_0 | x) P(x) dx}, \quad (3.9)$$

où $P(x_j)$ est la probabilité a priori que la mutation causale soit située à une position $x_j \in X$. Étudions maintenant la façon d'estimer $P(\Phi, H_0 | x_j)$. Pour cela, introduisons la notation T_x afin de représenter un arbre correspondant à la généalogie du marqueur x ; notons que la différence majeure entre cette méthode et les deux présentées précédemment est qu'ici on ne considère plus des graphes de recombinaison ancestraux, mais bien des arbres correspondant à la généalogie d'un unique marqueur. On obtient

alors

$$\begin{aligned} P(\Phi, H_0 | x_j) &= \int P(\Phi, H_0 | x_j, T_{x_j}) P(T_{x_j} | x_j) dT_{x_j} \\ &= \int P(\Phi | x_j, T_{x_j}) P(H_0 | \Phi, x_j, T_{x_j}) P(T_{x_j} | x_j) dT_{x_j}, \end{aligned} \quad (3.10)$$

où l'intégrale est évaluée sur l'espace des arbres à n feuilles. En supposant que le statut de cas ou de témoin des séquences n'affecte pas la généalogie du marqueur de la mutation causale, c'est-à-dire qu'il n'y a pas de sélection naturelle, il est possible de faire l'approximation suivante : $P(T_{x_j} | x_j) \approx P(T_{x_j})$. De plus, une supposition du modèle est que le marqueur sur lequel le TIM est apparu n'est pas un des L marqueurs des séquences de l'échantillon, ainsi l'approximation $P(H_0 | \Phi, x_j, T_{x_j}) \approx P(H_0 | T_{x_j})$ est valide. L'équation 3.10 devient donc

$$P(\Phi, H_0 | x_j) \approx \int P(\Phi | x_j, T_{x_j}) P(H_0 | T_{x_j}) P(T_{x_j}) dT_{x_j},$$

et puisque $P(H_0 | T_{x_j}) P(T_{x_j}) = P(T_{x_j} | H_0) P(H_0)$, on obtient

$$\begin{aligned} P(\Phi, H_0 | x_j) &\approx \int P(\Phi | x_j, T_{x_j}) P(T_{x_j} | H_0) P(H_0) dT_{x_j} \\ &\propto \int P(\Phi | x_j, T_{x_j}) P(T_{x_j} | H_0) dT_{x_j} \\ &\approx \frac{1}{M} \sum_{m=1}^M P(\Phi | x_j, T_{x_j}^{(m)}), \end{aligned} \quad (3.11)$$

où $T_{x_j}^{(m)}$ est généré selon la distribution a posteriori $P(T_{x_j} | H_0)$. La simulation d'arbre T_x tel que $T_x \sim T_x | H_0$ est exécutée à l'aide d'une chaîne de Markov Monte Carlo basée sur un modèle de coalescence avec recombinaison. Les détails de cette simulation sont disponibles dans la littérature (Zöllner et Pritchard, 2004). En remplaçant $P(\Phi, H_0 | x_j)$, dans l'équation 3.9 de la distribution a posteriori, par l'approximation obtenue à l'équation 3.11, on obtient finalement l'estimation suivante de la densité a posteriori d'intérêt :

$$P(x_j | \Phi, H_0) \approx \frac{(1/M) \sum_{m=1}^M P(\Phi | x_j, T_{x_j}^{(m)}) P(x_j)}{\sum_{i=1}^k \left((1/M) \sum_{m=1}^M P(\Phi | x_i, T_x^{(m)}) P(x_i) \right)}. \quad (3.12)$$

3.4.2 Calcul de la probabilité du phénotype

Nous allons maintenant analyser brièvement la façon de calculer la probabilité $P(\Phi|x, T_x)$; les détails de ce calcul seront présentés à la section 4.2.2 du chapitre suivant. Afin de calculer cette probabilité, il faut tout d'abord inférer un allèle au marqueur x ; la technique d'inférence utilisée ici est la même que celle dans la méthode Margarita. En effet, il suffit d'ajouter la mutation causale sur une des branches de la généalogie T_x et d'observer quelles séquences de notre échantillon en hériteront ; ceci nous donne une répartition des séquences en deux groupes : porteuses et non-porteuses du TIM. Posons B_{T_x} pour l'ensemble des branches de l'arbre T_x ; la probabilité recherchée peut alors s'écrire comme une somme sur toutes les branches sur lesquelles la mutation aurait pu survenir :

$$P(\Phi|x, T_x) = \sum_{w \in B_{T_x}} \left(f_0^{n_c^{np}} \cdot (1 - f_0^{n_t^{np}}) \cdot f_1^{n_c^p} \cdot (1 - f_1^{n_t^p}) \cdot P(w|T_x) \right), \quad (3.13)$$

$$\text{où } \begin{cases} n_c^{np} &= \text{nombre de cas au statut non-porteur} \\ n_t^{np} &= \text{nombre de témoins au statut non-porteur} \\ n_c^p &= \text{nombre de cas au statut porteur} \\ n_t^p &= \text{nombre de témoins au statut porteur,} \end{cases}$$

et $P(w|T_x)$ correspond à la probabilité qu'une mutation soit apparue sur la branche w de l'arbre T_x . Rappelons de plus que $F = (f_0, f_1)$ est la fonction de pénétrance pour la maladie influencée par le TIM.

L'approximation 3.12 se calcule donc aisément et l'estimateur de la position du TIM est le mode de la distribution a posteriori, c'est-à-dire la position $x \in X$ ayant la plus grande probabilité a posteriori.

3.4.3 Exemple de résultat

La figure 3.11 illustre les résultats obtenus avec la méthode LATG afin de détecter la mutation influençant la maladie de la fibrose kystique (Zöllner and Pritchard, 2005). Notons cependant que la courbe illustrée sur ce graphique n'est pas la distribution a

posteriori de l'équation 3.12, mais plutôt une vraisemblance a posteriori moyenne définie de la façon suivante :

$$L(x_j|\Phi, H_0) \approx \frac{1}{M} \sum_{m=1}^M P(\Phi|x_j, T_{x_j}^{(m)}). \quad (3.14)$$

Cette dernière équation est donc équivalente au numérateur de l'équation 3.12 divisé par la probabilité a posteriori $P(x_j)$.

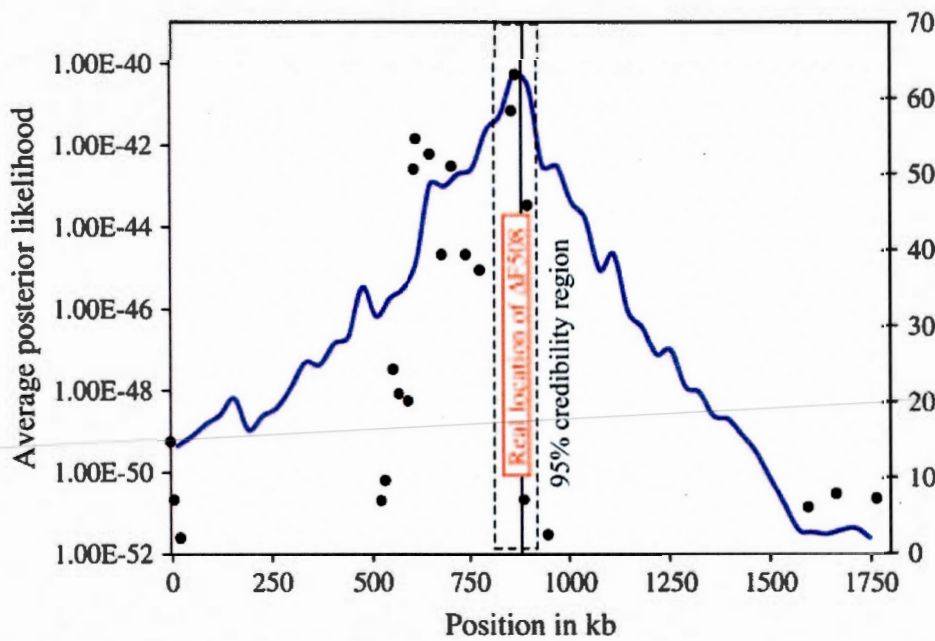


Figure 3.11 Graphique représentant la distribution de la vraisemblance a posteriori moyenne (ligne bleue) obtenue avec LATAG pour les données de la fibrose kystique. Nous pouvons y observer que le maximum de cette distribution est situé très près de la vraie position de la mutation causale (ligne verticale noire), l'estimation de la méthode LATAG est donc très bonne pour cet exemple (tiré de Zöllner et Pritchard, 2005).

CHAPITRE IV

UNE NOUVELLE MÉTHODE DE CARTOGRAPHIE GÉNÉTIQUE FINE

Nous avons présenté au chapitre précédent trois méthodes de cartographie génétique fine qui utilisaient différentes stratégies afin d'aborder les problèmes liés à l'estimation de la position d'une mutation influençant une maladie génétique. En effet, la façon de construire une généalogie et d'inférer un statut de mutant ou de non-mutant à chaque séquence d'un échantillon est différente pour chacune de ces méthodes. Par exemple, la méthode MapARG utilise le graphe de recombinaison ancestral afin de construire des généalogies à partir d'un échantillon de séquences génétiques tandis que la méthode Margarita emploie un algorithme heuristique basé sur la théorie de la coalescence avec recombinaison. Pour sa part, la méthode LATAG emploie des chaînes de Markov Monte Carlo pour construire les arbres partiels des marqueurs contenus dans les séquences génétiques d'un échantillon plutôt que de construire des graphes représentant la généalogie des séquences entières. De plus, la façon d'inférer un allèle mutant («1») ou primitif («0») à un marqueur donné pour chacune des séquences d'un échantillon est identique pour les méthodes Margarita et LATAG; en effet, la stratégie employée par ces méthodes est d'ajouter la mutation causale sur une des branches de l'arbre partiel du marqueur donné. De son côté, la méthode MapARG utilise une façon très différente de faire cette inférence; en effet, celle-ci est faite avant la simulation de la généalogie, et ce, par un algorithme EM. Finalement, l'estimateur de la position d'un TIM est différent pour chaque méthode. L'estimateur employé par la méthode Margarita est la position sur la séquence génétique ayant l'association la plus significative avec le phénotype, tandis que

celui utilisé par la méthode MapARG est un estimateur à maximum de vraisemblance. Pour sa part, la méthode LATAG estime la loi a posteriori d'un ensemble de positions possibles pour le TIM et l'estimateur de l'emplacement du TIM est le mode de cette distribution.

Le but de ce chapitre est de présenter une nouvelle méthode, appelé DMap, qui est constitué des meilleurs éléments de chacune des méthodes présentées dans le chapitre précédent. Cette méthode a été développée initialement dans le but d'améliorer l'étape d'inférence de la méthode MapARG pour des maladies présentant des pénétrances incomplètes et/ou des phénocopies.

4.1 Idée générale

Les échantillons de données considérés par la méthode DMap sont les mêmes que ceux considérés par les trois méthodes présentées au chapitre précédent et décrits à la section 3.1 ; ils contiennent de courtes séquences génétiques, formées de SNPs, dont on connaît le phénotype pour une maladie génétique d'intérêt, et l'unique mutation (TIM) influençant cette maladie génétique est située sur la portion d'ADN considérée. De plus, nous allons supposer que la fréquence de la maladie d'intérêt dans la population de séquences génétiques d'où proviennent nos échantillons ainsi que sa fonction de pénétrance $F = (f_0, f_1)$ sont connues. Notons qu'il n'est pas très réaliste de supposer la fonction de pénétrance connue, néanmoins, ceci nous permettra de développer une méthode performante dans ce contexte et une fois cette méthode développée, il sera plus facile de la modifier afin d'éliminer cette supposition contraignante ; nous aborderons ce sujet à la section 5.4 du prochain chapitre.

L'objectif de la méthode DMap est de calculer la vraisemblance de plusieurs positions potentielles pour l'emplacement du TIM ; l'estimation de cet emplacement sera la position ayant la vraisemblance la plus élevée. Les sections suivantes décrivent en détail la façon d'évaluer la vraisemblance d'une position donnée.

4.2 Le modèle

4.2.1 Calcul de la fonction de vraisemblance

Nos données sont formées par un échantillon contenant n séquences génétiques formées de L marqueurs d'allèle et de position connus. Dénotons $\Phi = (\phi_1, \phi_2, \dots, \phi_n)$ le vecteur contenant le phénotype de chacune des n séquences de nos données et H_0 l'ensemble contenant les séquences génétiques de notre échantillon. L'objectif de DMap étant de calculer la vraisemblance de plusieurs positions possibles pour le TIM, nous allons définir $Y = \{y_1, y_2, \dots, y_Z\}$, l'ensemble contenant les Z positions à évaluer, $x_{yz}, z \in \{1, 2, \dots, Z\}$, un marqueur situé à la position y_z sur une séquence génétique et c_T la position inconnue du TIM. Similairement à la méthode MapARG, nous pouvons écrire la fonction de vraisemblance pour c_T de la façon suivante :

$$\begin{aligned} L(c_T) &= P(H_0, \Phi | c_T) \\ &= \int P(H_0, \Phi | G, c_T) \cdot P(G | c_T) \cdot dG \\ &= \int P(H_0 | \Phi, G, c_T) \cdot P(\Phi | G, c_T) \cdot P(G | c_T) \cdot dG, \end{aligned} \quad (4.1)$$

où G représente une généalogie. En utilisant une méthode d'échantillonnage pondéré avec la distribution proposée $Q(\cdot)$ de Fearnhead et Donnelly pour générer des généalogies consistantes avec les données contenues dans H_0 , on peut approximer l'équation 4.1 par

$$L(c_T) \approx \frac{1}{M} \sum_{i=1}^M \frac{P(H_0 | \Phi, G^{(i)}, c_T) \cdot P(\Phi | G^{(i)}, c_T) \cdot P(G^{(i)} | c_T)}{Q(G^{(i)})}. \quad (4.2)$$

Il est important de remarquer que contrairement à la méthode MapARG, nous supposons que les séquences contenues dans H_0 ne contiennent pas de marqueur représentant la mutation causale, ainsi les généalogies générées sont indépendantes de la position de la mutation causale et donc $P(G^{(i)} | c_T) = P(G^{(i)})$. Cette remarque nous permet aussi de supposer que les données contenues dans H_0 sont indépendantes du phénotype lié à la maladie d'intérêt ainsi que de la position de la mutation, ce qui implique que $P(H_0 | \Phi, G^{(i)}, c_T) = P(H_0 | G^{(i)})$. De plus, les généalogies $G^{(i)}$ étant consistantes avec les données, nous obtenons $P(H_0 | G^{(i)}) = 1, \forall i$. L'équation 4.2 peut donc se réécrire

comme :

$$L(c_T) \approx \frac{1}{M} \sum_{i=1}^M \frac{P(\Phi|G^{(i)}, c_T) \cdot P(G^{(i)})}{Q(G^{(i)})}. \quad (4.3)$$

Premièrement, afin de calculer les probabilités $P(G)$ et $Q(G)$, considérons une méthode qui est exactement la même que celle employée dans la méthode MapARG (voir les sections 3.2.2 (page 30) et 3.2.3 (page 39)). L'équation 4.3 peut donc se réécrire de la façon suivante :

$$\begin{aligned} L(c_T) &\approx \frac{1}{M} \sum_{i=1}^M \left(P(\Phi|G^{(i)}, c_T) \cdot \prod_{\tau=0}^{\tau^*-1} \left[\frac{P(H_\tau^{(i)}|H_{\tau+1}^{(i)})}{Q(H_{\tau+1}^{(i)}|H_\tau^{(i)})} \right] \right) \\ &= \frac{1}{M} \sum_{i=1}^M \left(P(\Phi|G^{(i)}, c_T) \cdot \prod_{\tau=0}^{\tau^*-1} \left[\frac{P(H_\tau^{(i)}|H_{\tau+1}^{(i)})}{P(H_\tau^{(i)}|H_{\tau+1}^{(i)}) \cdot \pi(H_{\tau+1}^{(i)})/\pi(H_\tau^{(i)})} \right] \right) \\ &= \frac{1}{M} \sum_{i=1}^M \left(P(\Phi|G^{(i)}, c_T) \cdot \prod_{\tau=0}^{\tau^*-1} \left[\frac{\pi(H_\tau^{(i)})}{\pi(H_{\tau+1}^{(i)})} \right] \right). \end{aligned} \quad (4.4)$$

Deuxièmement, lors de la description de la méthode Margarita au chapitre précédent (section 3.3.2), il a été mentionné que tous les graphes de recombinaison ancestraux générés étaient considérés équiprobables. En utilisant la même supposition dans la méthode DMap, nous obtenons que les probabilités $P(G)$ et $Q(G)$ n'ont plus d'impact sur la vraisemblance. L'équation 4.3 prend donc la forme suivante :

$$\begin{aligned} L(c_T) &\approx \frac{1}{M} \sum_{i=1}^M \frac{P(\Phi|G^{(i)}, c_T) \cdot P(G)}{Q(G)} \\ &\approx \frac{1}{M} \frac{P(G)}{Q(G)} \cdot \sum_{i=1}^M P(\Phi|G^{(i)}, c_T) \\ &\propto \frac{1}{M} \cdot \sum_{i=1}^M P(\Phi|G^{(i)}, c_T) \end{aligned} \quad (4.5)$$

Dans le chapitre suivant, nous allons analyser et comparer les résultats obtenus en considérant que les poids des graphes simulés sont différents (équation 4.4) et en les considérant équivalents (équation 4.5).

4.2.2 Calcul de la probabilité du phénotype

Nous allons maintenant nous intéresser à la façon de calculer la probabilité d'observer les phénotypes de chaque séquence de notre échantillon, sachant la généalogie de cet échantillon et l'emplacement de la mutation causale, c'est-à-dire la probabilité $P(\Phi|G, c_T)$. Notons que la stratégie pour calculer cette probabilité est inspirée de la méthode LATAG (section 3.4). Pour évaluer cette probabilité, nous devons tout d'abord assigner à chaque séquence de l'échantillon un statut de porteuse ou de non-porteuse du TIM. Le tableau 4.1 présente les quatre catégories possibles pour chaque séquence une fois cette assignation faite ainsi que la probabilité de chacune de celles-ci. Rappelons que f_0 représente la probabilité qu'une séquence génétique provienne d'un individu malade sachant qu'elle n'est pas porteuse du TIM, et donc que $(1 - f_0)$ représente la probabilité qu'une séquence génétique provienne d'un individu non-malade sachant qu'elle n'est pas porteuse du TIM.

Catégories			
Langage courant	Langage mathématique	Notations	$P(\phi T)$
Cas et porteuse	$\phi = 1, T = 1$	cp	f_1
Cas et non-porteuse	$\phi = 1, T = 0$	cnp	f_0
Témoin et porteuse	$\phi = 0, T = 1$	tp	$1 - f_1$
Témoin et non-porteuse	$\phi = 0, T = 0$	tnp	$1 - f_0$

Tableau 4.1 Tableau résumant les 4 catégories possibles pour une séquence. La variable ϕ représente le phénotype de la séquence, tel que $\phi = 1$ correspond à un cas (une séquence provenant d'un individu malade) et $\phi = 0$ correspond à un témoin. La variable T correspond au nombre d'allèles mutants que possède la séquence, elle peut donc prendre les valeurs de 0 ou de 1.

La méthode d'inférence employée afin d'assigner une catégorie à chaque séquence est basée sur la méthode Margarita (section 4.4). Il faut premièrement ajouter un marqueur à la position c_T sur chacune des séquences de la généalogie G ; pour l'instant, l'allèle du marqueur x_{c_T} est inconnu pour toutes les séquences. Par la suite, nous pourrions

extraire de la généalogie G , l'arbre partiel du marqueur x_{c_T} , que nous noterons A_{c_T} . La deuxième étape consiste à ajouter la mutation causale sur l'une des branches de l'arbre A_{c_T} . Nous pouvons par la suite inférer un allèle muté («1») au marqueur x_{c_T} de toutes les séquences de l'arbre A_{c_T} qui ont hérité de la mutation causale et un allèle non-muté («0») à toutes les autres séquences. En observant les séquences résultantes, nous pouvons aisément calculer le nombre de séquences dans chaque catégorie; nous noterons n_a le nombre de séquences présentes dans la catégorie a . Ainsi, en supposant que la mutation causale est apparue sur la branche b de l'arbre A_{c_T} , nous obtenons la probabilité conditionnelle suivante :

$$\begin{aligned}
 P(\Phi|G, c_T, b) &= \prod_{j=1}^n P(\phi_j|G, c_T, b) \\
 &= \prod_{j=1}^n f_{T_j}^{(\phi_j)} (1 - f_{T_j})^{(1-\phi_j)} \\
 &= f_0^{n_{cp}} \cdot (1 - f_0)^{n_{tp}} \cdot f_1^{n_{cp}} \cdot (1 - f_1)^{n_{tp}}, \quad (4.6)
 \end{aligned}$$

où T_j correspond au nombre d'allèles mutants que possède la séquence j et ϕ_j correspond au phénotype de la séquence j ; les détails de cette notation ont été présentés dans le tableau 4.1. Il est maintenant important de remarquer que nous ne connaissons pas la branche de l'arbre partiel sur laquelle la mutation causale s'est produite. Cependant, en posant $B_{A_{c_T}}$ l'ensemble des branches de l'arbre A_{c_T} , il est possible de sommer sur toutes les branches de l'arbre et d'obtenir l'équation suivante :

$$\begin{aligned}
 P(\Phi|G, c_T) &= \sum_{b \in B_{A_{c_T}}} P(\Phi|G, c_T, b) \cdot P(b|G, c_T) \\
 &= \sum_{b \in B_{A_{c_T}}} f_0^{n_{cp}} \cdot (1 - f_0)^{n_{tp}} \cdot f_1^{n_{cp}} \cdot (1 - f_1)^{n_{tp}} \cdot P(b|G, c_T). \quad (4.7)
 \end{aligned}$$

La probabilité $P(b|G, c_T)$ se calcule aisément à l'aide d'une propriété du processus de coalescence avec mutation énoncée à la section 2.3. Cette propriété stipulait qu'une fois que le nombre de mutations apparaissant sur une branche était connu, les temps auxquels apparaissaient ces mutations étaient aléatoires. Ceci nous permet d'affirmer que la position de la mutation que l'on ajoute sur l'arbre A_{c_T} est aléatoire et donc que la probabilité de l'ajouter sur une branche donnée est proportionnelle à la longueur de

cette branche. On obtient donc :

$$P(b|G, c_T) = \frac{|b|}{\sum_{w \in B_{Ac_T}} |w|}, \quad (4.8)$$

où $|b|$ représente la longueur de la branche b . Nous pouvons interpréter cette probabilité de la façon suivante : plus une branche b est longue, plus elle a un poids élevé dans le calcul de la probabilité du phénotype. Il serait cependant intéressant que cette probabilité prenne aussi en considération la répartition des séquences génétiques dans les catégories porteuses et non-porteuses du TIM obtenue suite à l'ajout du TIM sur une branche b . En effet, si en posant le TIM sur une branche b , le nombre de séquences héritant de la mutation dans notre échantillon est probable, c'est-à-dire que la fréquence de la mutation observée dans notre échantillon est similaire à la fréquence théorique de la mutation, nous aimerions alors que cette branche b ait un poids important dans l'équation 4.7. Nous avons donc choisi de modifier la probabilité $P(b|G, c_T)$ afin que son poids dans l'équation 4.7 soit représentatif de la vraisemblance de la fréquence de la mutation observée dans notre échantillon suite à l'ajout du TIM sur la branche b en plus d'être représentatif de la longueur de la branche b . Afin d'expliquer la modification que nous lui avons apportée, nous allons introduire quelques notations. Nous allons noter n le nombre de séquences dans notre échantillon et $n_m^{(b)}$ le nombre de séquences de notre échantillon qui sont dans la catégorie porteuse de la mutation suite à son ajout sur la branche b . Posons de plus N_m la variable aléatoire représentant le nombre de séquences porteuses de la mutation causale dans notre échantillon et μ_e la fréquence théorique de la mutation dans notre échantillon. Nous obtenons ainsi que :

$$N_m \sim \text{Bin}(n, \mu_e),$$

et l'approximation suivante de l'équation 4.8 :

$$\begin{aligned} P(b|G, c_T) &= P(b|G, c_T, \mu_e) \\ &\propto \frac{|b|}{\sum_{w \in B_{Ac_T}} |w|} \cdot P(N_m = n_m^{(b)}). \end{aligned} \quad (4.9)$$

Notons que la modification apportée à la probabilité $P(b|G, c_T)$ est basée sur la connaissance de μ_e , la fréquence de la mutation dans notre échantillon. Nous allons maintenant

présenter le calcul permettant d'obtenir cette fréquence. Pour cela, il faut tout d'abord obtenir la fréquence de la mutation dans notre population que nous noterons μ_p . Cette fréquence s'obtient facilement, car nous avons supposé à la section 4.1 la fréquence de la maladie (notée t_m) ainsi que la fonction de pénétrance $F = (f_0, f_1)$ connues. On obtient donc à l'aide de la formule des probabilités totales :

$$\begin{aligned} P(\phi = 1) &= P(\phi = 1|T = 1) \cdot P(T = 1) + P(\phi = 1|T = 0) \cdot P(T = 0) \\ &= P(\phi = 1|T = 1) \cdot P(T = 1) + P(\phi = 1|T = 0) \cdot (1 - P(T = 1)) \\ &= P(T = 1) \cdot [P(\phi = 1|T = 1) - P(\phi = 1|T = 0)] + P(\phi = 1|T = 0) \end{aligned}$$

En isolant $P(T = 1)$ dans l'égalité précédente, nous obtenons

$$P(T = 1) = \frac{P(\phi = 1) - P(\phi = 1|T = 0)}{P(\phi = 1|T = 1) - P(\phi = 1|T = 0)} \quad (4.10)$$

où $P(\phi = 1)$ correspond à la probabilité qu'une séquence génétique provienne d'un individu malade et donc à la fréquence de la maladie (t_m) et $P(T = 1)$ correspond à la probabilité qu'une séquence génétique soit porteuse du TIM et donc à la fréquence de la mutation (μ_p). De plus, la probabilité $P(\phi = 1|T = i), i = \{0, 1\}$, représente la probabilité qu'une séquence génétique provienne d'un individu malade sachant qu'elle est non-porteuse ($T = 0$) ou porteuse ($T = 1$) du TIM, elle correspond donc à f_0 et f_1 , respectivement. Ainsi, l'équation 4.10 peut se réécrire de la façon suivante :

$$P(T = 1) = \mu_p = \frac{t_m - f_0}{f_1 - f_0}.$$

Maintenant que nous avons montré la façon d'obtenir la fréquence de la mutation dans une population, nous allons étudier la façon d'obtenir μ_e à partir de celle-ci. Nous pouvons premièrement remarquer que si nous utilisons un échantillon aléatoire de notre population, alors $\mu_e = \mu_p$. Cependant, puisque nous travaillons avec des maladies peu fréquentes, il est commun de travailler avec des échantillons formés d'un même nombre de cas et de témoins; ces échantillons étant aléatoires stratifiés, nous devons trouver une méthode afin de calculer μ_e pour ceux-ci. Pour cela, il suffit de constater que :

$$\mu_e = \frac{n_c \cdot P(T = 1|\phi = 1) + n_t \cdot P(T = 1|\phi = 0)}{n}, \quad (4.11)$$

où n_c et n_t représentent respectivement le nombre de cas et de témoins dans notre échantillon de taille n , et $P(T = 1|\phi = i)$, $i = \{0, 1\}$, représente la probabilité qu'une séquence génétique soit porteuse du TIM, sachant qu'elle provient d'un individu malade ($\phi = 1$) ou non-malade ($\phi = 0$). Les probabilités conditionnelles de l'équation 4.11 s'obtiennent facilement grâce au théorème de Bayes :

$$\begin{aligned} P(T = 1|\phi = 1) &= \frac{P(\phi = 1|T = 1) \cdot P(T = 1)}{P(\phi = 1)} \\ &= \frac{f_1 \cdot \mu_p}{t_m}, \end{aligned} \quad (4.12)$$

$$\begin{aligned} P(T = 1|\phi = 0) &= \frac{P(\phi = 0|T = 1) \cdot P(T = 1)}{P(\phi = 0)} \\ &= \frac{(1 - f_1) \cdot \mu_p}{1 - t_m}. \end{aligned} \quad (4.13)$$

En substituant les probabilités conditionnelles de l'équation 4.11 par les expressions obtenues aux équations 4.12 et 4.13, nous obtenons finalement :

$$\mu_e = \left(n_c \cdot \frac{f_1 \cdot \mu_p}{t_m} + n_t \cdot \frac{(1 - f_1) \cdot \mu_p}{1 - t_m} \right) / n.$$

Le chapitre suivant présente les résultats obtenus en utilisant les deux formes de la probabilité $P(b|G, c_T)$: sa forme exacte (équation 4.8), qui ne tient pas compte de la fréquence de la mutation, et sa forme approximative (équation 4.9) qui, elle, n'ignore pas des informations existantes.

La figure 4.1 illustre un exemple récapitulatif de la méthode où la position $\pi_z \in \pi$, dont on veut calculer la vraisemblance, est située entre le deuxième et le troisième marqueur des séquences de notre échantillon. Nous pouvons y voir la façon d'ajouter un marqueur à cette position sur toutes les séquences de la généalogie et ensuite d'extraire l'arbre partiel de ce nouveau marqueur. Les branches de l'arbre partiel obtenu sont identifiées par des lettres de A à F ; chacune de ces branches a une certaine probabilité de recevoir la mutation causale. Finalement, l'arbre partiel obtenu suite à l'ajout du TIM sur la branche E est illustré ; nous pouvons ainsi facilement voir les séquences qui héritent de la mutation causale.

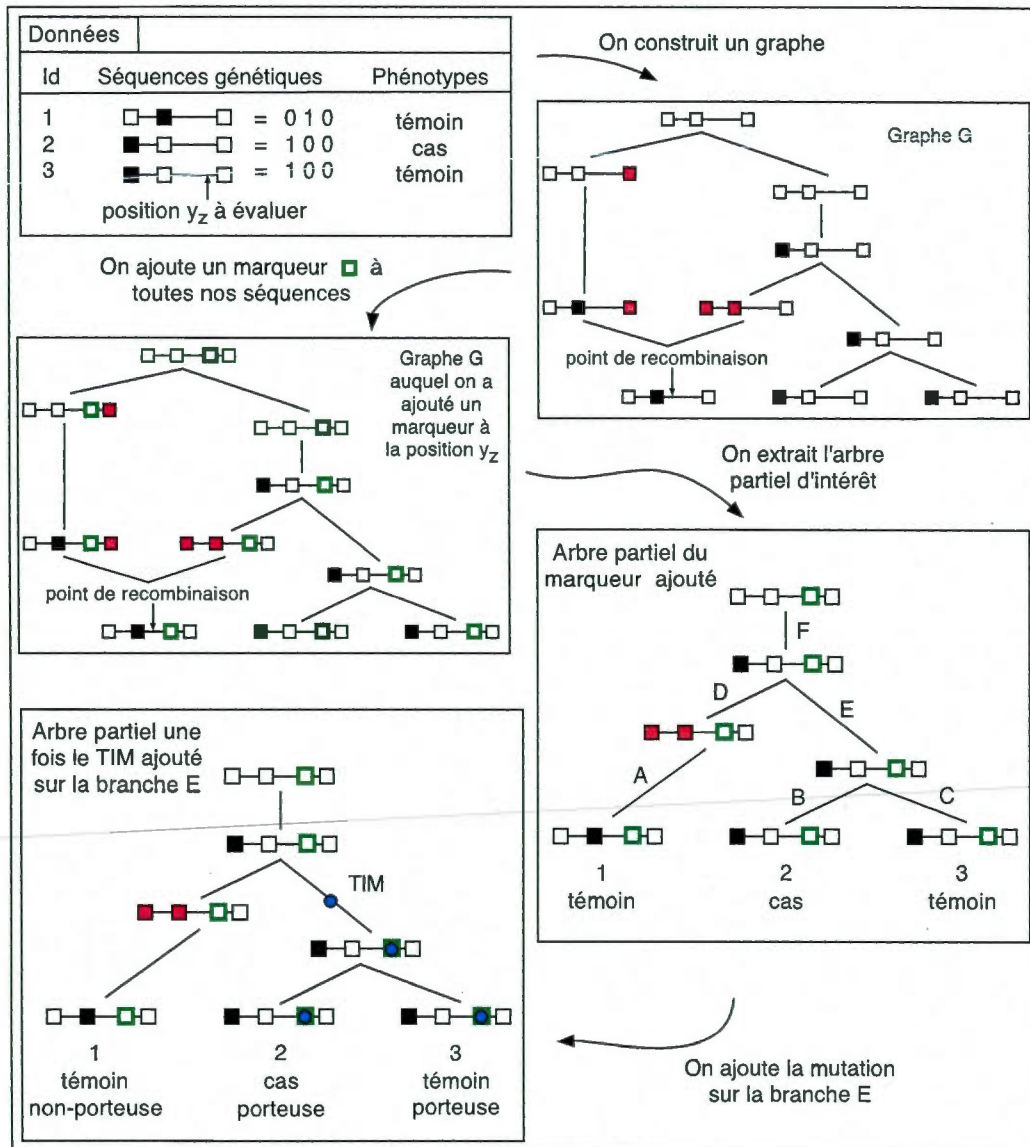


Figure 4.1 Illustration des étapes afin d'obtenir une répartition des séquences contenues dans nos données dans les 4 catégories d'intérêt. L'encadré en haut à gauche illustre les données ; elles sont formées de 3 séquences génétiques accompagnées de leur phénotype. L'étape suivante est de générer un graphe consistant avec les données. Par la suite, un marqueur, représenté par un carré vert, est ajouté à la position y_z sur chaque séquence du graphe. L'encadré en bas à droite illustre l'arbre partiel du marqueur ajouté. Cet arbre contient 6 branches, nous devons calculer la probabilité du phénotype conditionnellement à l'ajout du TIM sur chacune de celles-ci. Finalement, le dernier encadré illustre l'arbre partiel obtenu une fois que le TIM (cercle bleu) a été ajouté sur la branche E ; les séquences 2 et 3 héritent du TIM.

Le tableau 4.2 présente, pour chacune des branches de l'arbre partiel illustré à la figure 4.1, la répartition des séquences de notre échantillon dans les quatre catégories d'intérêt, suite à l'ajout du TIM sur la branche considérée.

Branche sur laquelle on ajoute le TIM	Séquences qui héritent du TIM	n_{cp}	n_{cnp}	n_{tp}	n_{tnp}
A	1	0	1	1	1
B	2	1	0	0	2
C	3	0	1	1	1
D	1	0	1	1	1
E	2,3	1	0	1	1
F	1,2,3	1	0	2	0

Tableau 4.2 Tableau présentant les séquences qui héritent de la mutation causale lors de l'ajout du TIM sur chacune des branches de l'arbre partiel illustré à la figure 4.1. Nous pouvons observer de plus le nombre de séquences dans chacune des 4 catégories possibles pour les séquences, soit de gauche à droite, cas et porteuse, cas et non-porteuse, témoin et porteuse et finalement témoin et non porteuse.

4.3 Algorithme et défi computationnel

4.3.1 Algorithme de DMap

Nous allons tout d'abord présenter l'algorithme sous-jacent à la méthode DMap afin de pouvoir discuter par la suite des défis computationnels qu'il entraîne.

1. Choisir un ensemble $Y = (y_1, \dots, y_Z)$ de valeurs possibles et à évaluer pour la position c_T du TIM.
2. Poser H_0 pour l'ensemble contenant les n séquences de notre échantillon ainsi que Φ pour le vecteur contenant le phénotype de chacune des séquences.
3. Pour $i = 1, \dots, M$ (boucle sur les ARGs) faire

- (a) Construire un graphe de recombinaison ancestral $G^{(i)}$ à l'aide de la distribution proposée de Fearnhead et Donnelly en suivant les étapes suivantes :
- i. Poser $\tau = 0$.
 - ii. Tant que $|H_\tau^{(i)}| \neq 1$, c'est-à-dire tant que nous n'avons pas atteint le MRCA, faire
 - Générer le temps $t_\tau^{(i)}$ avant le prochain événement tel que $t_\tau^{(i)} \sim \exp(\frac{n(n-1+\alpha\theta+\beta\rho)}{2})$, où $n = |H_\tau^{(i)}|$, c'est-à-dire le nombre de séquences présentes à l'état $H_\tau^{(i)}$.
 - Calculer $Q(H_{\tau+1}^{(i)}|H_\tau^{(i)})$ à l'aide de l'équation 3.8 pour tous les états $H_{\tau+1}^{(i)}$ que l'on peut atteindre de l'état $H_\tau^{(i)}$, i.e tel que $P(H_\tau^{(i)}|H_{\tau+1}^{(i)}) > 0$.
 - Choisir le prochain état $H_{\tau+1}^{(i)}$ de l'ARG proportionnellement à sa pondération.
 - Poser $\tau = \tau + 1$.
- (b) Calculer la quotient $R^{(i)} = P(G^{(i)})/Q(G^{(i)}) = \prod_{\tau=0}^{\tau^*-1} \pi(H_\tau^{(i)})/\pi(H_{\tau+1}^{(i)})$.
- (c) Pour $z = 1, \dots, Z$ (boucle sur les positions dont on veut calculer la vraisemblance) faire
- i. Ajouter un marqueur à la position y_z sur chaque séquence du graphe $G^{(i)}$.
 - ii. Extraire l'arbre partiel A_{y_z} et construire l'ensemble $B_{A_{y_z}} = \{b_1, \dots, b_{B_z}\}$ contenant les branches de A_{y_z} .
 - iii. Pour chaque $b \in B_{A_{y_z}}$, calculer $H_b = P(\Phi|G^{(i)}, y_z, b) \cdot P(b|G^{(i)}, y_z)$ à l'aide des équations 4.6 et 4.8.
 - iv. Calculer $P(\Phi|G^{(i)}, y_z) = \sum_{b \in B_{A_{y_z}}} H_b$.
4. Pour $z = 1, \dots, Z$, calculer $L(y_z) = \frac{1}{M} \sum_{i=1}^M R^{(i)} \cdot P(\Phi|G^{(i)}, y_z)$.
5. L'estimateur de la position c_T est $y_z^* \in Y$ tel que $\max\{L(y_1), \dots, L(y_Z)\} = L(y_z^*)$.

4.3.2 Défis computationnels de DMap

Le défi computationnel majeur de l'algorithme présenté à la sous-section précédente est de trouver une méthode efficace permettant d'exécuter les étapes 3(c)i à 3(c)iii, c'est-à-dire permettant d'ajouter un marqueur à une position précise sur chacune des séquences d'un ARG (3(c)i), d'extraire de ce même ARG l'arbre partiel du marqueur ajouté (3(c)ii) et de pouvoir ensuite identifier rapidement les séquences de notre échantillon qui hériteraient du TIM suite à son ajout sur une des branches de l'arbre partiel, et ce, pour chaque branche de l'arbre (3(c)iii). Tout d'abord, il est important de remarquer qu'afin de pouvoir extraire un arbre partiel d'un ARG, il faut connaître les informations contenues dans cet ARG ; il est donc nécessaire de garder en mémoire les ARGs générés. La première étape est donc de trouver une façon simple et efficace de condenser l'information d'intérêt contenue dans un ARG afin de pouvoir par la suite l'emmagasiner dans un objet informatique. La technique que nous avons employée est de premièrement assigner un numéro d'identification unique à toutes les séquences de l'ARG considéré. Il suffit par la suite de conserver en mémoire tous les événements transitoires qui ont eu lieu dans l'ARG, ainsi que les temps d'attente entre ceux-ci. Ceci se fait aisément en emmagasinant chaque événement dans une ligne d'une immense matrice. La figure 4.2 présente la notation utilisée afin de représenter chacun des quatre événements de transitions possibles, soit une coalescence entre deux séquences identiques, une coalescence entre deux séquences différentes, une mutation et une recombinaison. Notons que dans la méthode MapARG, il n'est pas nécessaire de garder en mémoire les ARGs générés, puisqu'une fois simulé, ils ne sont plus utiles : en effet, seules les probabilités de transitions sont importantes et celles-ci sont calculées lors de la construction de l'ARG.

Nous avons donc maintenant une façon efficace de conserver les informations pertinentes contenues dans un ARG. Avant de décrire la méthode permettant de relever le défi majeur mentionné ci-haut, rappelons quelques définitions provenant de la théorie des graphes afin d'en faciliter la présentation.

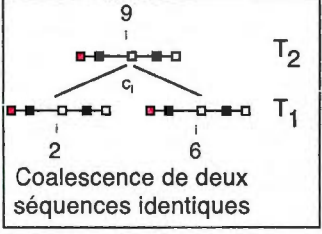
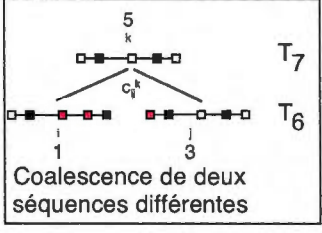
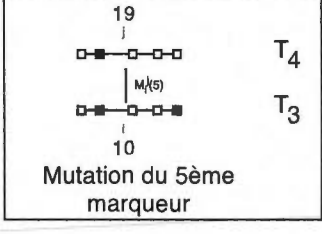
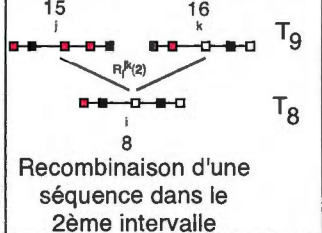
Événements	Notations
 <p>Coalescence de deux séquences identiques</p>	<p>Code de l'événement : 1</p> <p>Numéro de la 1ère séquence qui coalesce : 2</p> <p>Numéro de la 2ème séquence qui coalesce : 6</p> <p>Numéro de la séquence résultante : 9</p> <p>Temps d'attente pour que l'événement survienne : $T_2 - T_1$</p>
 <p>Coalescence de deux séquences différentes</p>	<p>Code de l'événement : 2</p> <p>Numéro de la 1ère séquence qui coalesce : 1</p> <p>Numéro de la 2ème séquence qui coalesce : 3</p> <p>Numéro de la séquence résultante : 5</p> <p>Temps d'attente pour que l'événement survienne : $T_7 - T_6$</p>
 <p>Mutation du 5ème marqueur</p>	<p>Code de l'événement : 3</p> <p>Numéro de la séquence qui mute : 10</p> <p>Numéro de la séquence résultante : 19</p> <p>Marqueur qui mute : 5</p> <p>Temps d'attente pour que l'événement survienne : $T_4 - T_3$</p>
 <p>Recombinaison d'une séquence dans le 2ème intervalle</p>	<p>Code de l'événement : 4</p> <p>Numéro de la séquence qui recombine : 8</p> <p>Numéro du parent de gauche : 15</p> <p>Numéro du parent de droite : 16</p> <p>Coordonnée exacte de la recombinaison : x</p> <p>Temps d'attente pour que l'événement survienne : $T_9 - T_8$</p>

Figure 4.2 Illustration de la notation utilisée afin de représenter d'une façon concise chaque événement possible d'un graphe. Les éléments de la colonne de gauche sont habituellement emmagasinés sur une ligne d'une matrice contenant l'information du graphe considéré.

- Un arbre est un graphe non-orienté, acyclique et connexe ; ceci signifie qu'un arbre de n sommets contient $n - 1$ arêtes non-orientées et que chaque sommet est au moins, de degré 1, c'est-à-dire que chaque sommet est au moins lié à un autre sommet par

une arête ;

- Les sommets de degré 1 sont appelés des feuilles et ceux de degré strictement supérieur à 1 sont appelés des nœuds internes ;
- Dans un graphe orienté, lorsqu'une arête pointe d'un sommet A à un sommet B, on dit que A est le parent de B. Les descendants d'un sommet C sont tous les sommets D tels qu'il existe un chemin orienté allant de A à D ;
- Un arbre enraciné est un graphe orienté, acyclique et connexe dans lequel un sommet est appelé la racine de l'arbre. Cette racine est l'unique sommet de l'arbre ne possédant pas de parent, tous les autres sommets ont un unique parent.

La figure 4.3 illustre les concepts dont nous venons de donner la définition. Il est maintenant possible de constater qu'un arbre partiel est en fait un arbre enraciné dont la racine représente l'ancêtre commun le plus récent (MRCA). Les feuilles de l'arbre sont représentées par les séquences génétiques de notre échantillon et les nœuds internes représentent toutes les autres séquences de notre généalogie.

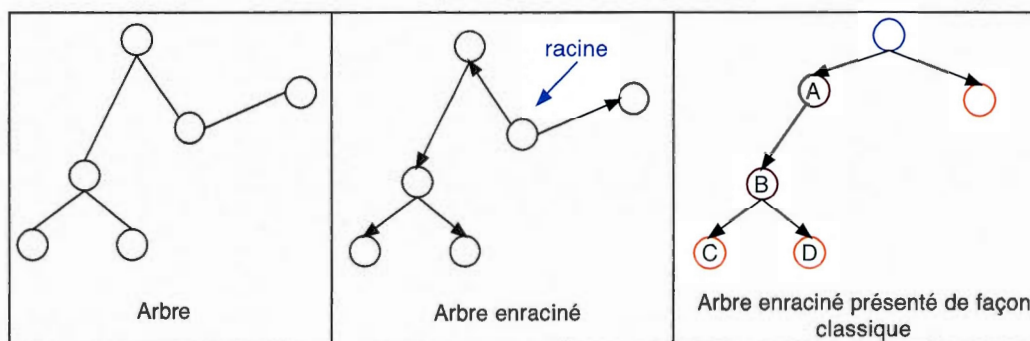


Figure 4.3 Illustration de quelques concepts de la théorie des graphes. La figure de gauche présente un arbre et celle du milieu un arbre enraciné. La figure de droite représente le même arbre enraciné que la figure du centre, mais d'une façon classique, c'est-à-dire avec la racine située tout en haut. De plus, les sommets oranges sont des feuilles et les sommets rouges sont des nœuds internes. Il est aussi possible d'observer que les feuilles C et D ont le même parent (le sommet B) et que les descendants du sommet A sont les sommets B, C et D.

Tous les éléments sont maintenant en place afin de décrire la méthode permettant d'exécuter efficacement les étapes 3(c)i à 3(c)iii de l'algorithme précédent. Notre méthode est basée sur le fait que les seules informations que nous devons connaître sur l'arbre A_{yz} afin de réaliser l'étape 3(c)iii sont contenues dans les trois ensembles suivants :

- l'ensemble $B_{A_{yz}} = \{b_1, \dots, b_{B_z}\}$ contenant les branches de l'arbre A_{yz} ;
- l'ensemble $T_{A_{yz}} = \{t_{b_1}, \dots, t_{b_{B_z}}\}$ contenant la longueur de chaque branche $b \in B_{A_{yz}}$;
- l'ensemble $D_{A_{yz}} = \{D_{b_1}, \dots, D_{b_{B_z}}\}$ où $D_{b_j} = (n_{cp}, n_{tp})$ est un vecteur contenant le nombre de séquences contenues dans les catégories «cas et porteuse» et «témoin et porteuse» suite à l'ajout du TIM sur la branche b_j .

Notons que la connaissance du vecteur D_{b_j} est équivalente à la connaissance du nombre de séquences dans chacune des catégories du tableau 4.1 suite à l'ajout du TIM sur la branche b_j . En effet, il est très facile de déduire le nombre de séquences dans les catégories «cas et non-porteuse» et «témoin et non-porteuse» à partir du vecteur D_{b_j} puisque nous connaissons le nombre total de cas (n_c) et de témoins (n_t) de notre échantillon.

Ainsi $n_{cnp} = n_c - n_{cp}$ et $n_{tnp} = n_t - n_{tp}$. Nous obtenons donc bien que les ensembles $B_{A_{yz}}$,

$T_{A_{yz}}$ et $D_{A_{yz}}$ nous permettent de calculer aisément la probabilité $H_b = P(\Phi|G^{(i)}, y_z, b) \cdot P(b|G^{(i)}, y_z)$. Notons de plus qu'il est possible d'associer chaque branche d'un arbre à une séquence. En effet, nous savons que chaque nœud (séquence), à l'exception de la racine (MRCA), d'un arbre possède un unique parent et donc une unique arête (branche) pointant sur lui. Nous pouvons donc associer chaque arête à l'unique sommet vers lequel elle pointe ; les séquences qui hériteront d'une mutation ajoutée sur une branche b_j sont les mêmes lorsque nous ajoutons directement la mutation sur le nœud (séquence) vers lequel l'arête b_j pointe. Nous obtenons donc que l'ensemble $N_{A_{yz}}$ contenant les séquences présentes dans l'arbre A_{yz} est équivalent à l'ensemble $B_{A_{yz}}$; nous travaillerons par la suite uniquement avec l'ensemble $N_{A_{yz}}$.

Une façon efficace d'obtenir les ensembles $N_{A_{yz}}$, $T_{A_{yz}}$ et $D_{A_{yz}}$ est d'employer une approche itérative. La première étape est d'initialiser les ensembles lorsque nous nous plaçons à la génération H_0 de la généalogie G . Par la suite, chaque itération correspond à une transition d'une génération à une autre dans le graphe G , et ce, du présent au

passé; il suffit donc de mettre à jour les ensembles à chaque événement de transition du graphe. L'algorithme qui suit permet de calculer itérativement les ensembles $N_{A_{yz}}$ et $D_{A_{yz}}$. Notons que la méthode afin d'obtenir l'ensemble $T_{A_{yz}}$ étant quelque peu complexe, nous ne présenterons pas l'algorithme sous-jacent à celle-ci; cependant le code écrit en C++ permettant l'obtention des trois ensembles d'intérêt est présenté à l'annexe B (page 119).

1. Construire un ensemble P contenant les numéros d'identification des feuilles du graphe G .
2. Pour chaque $x \in P$, construire un vecteur $D_x = [d_x^c, d_x^t]$ où d_x^c correspond au nombre de descendants de x (incluant le sommet x) qui sont des feuilles et qui ont un statut de cas, tandis que d_x^t correspond au nombre de descendants de x (incluant le sommet x) qui sont des feuilles et qui ont un statut de témoin. Poser D l'ensemble contenant les $|P|$ vecteurs D_x .
3. Pour $\tau = 1, \dots, \tau^*$
 - (a) Poser i l'événement de transition entre les générations $H_{\tau-1}$ et H_τ ;
 - (b) Si l'événement $i = (1, a, b, c, t)$ ou $(2, a, b, c, t)$, i.e l'événement est une coalescence, faire
 - i. Si $a \in P$ ou $b \in P$ faire
 - poser $P = P \cup c$ et construire $D_c = [d_c^c, d_c^t]$ où $d_c^c = 0$ et $d_c^t = 0$;
 - si $a \in P$, poser $d_c^c = d_a^c$ et $d_c^t = d_a^t$;
 - si $b \in P$, poser $d_c^c = d_c^c + d_b^c$ et $d_c^t = d_c^t + d_b^t$;
 - poser $D = D \cup D_c$.
 - (c) Si l'événement $i = (3, a, b, t)$, i.e l'événement est une mutation, faire
 - i. Si $a \in P$ faire
 - poser $P = P \cup b$ et construire $D_b = [d_b^c, d_b^t]$ où $d_b^c = d_a^c$ et $d_b^t = d_a^t$;
 - poser $D = D \cup D_b$.
 - (d) Si l'événement $i = (4, a, b, c, y, t)$, i.e l'événement est une recombinaison, faire
 - i. Si $a \in P$ faire

- Si $y < y_z$, alors $P = P \cup b$ et construire $D_b = [d_b^c, d_b^t]$ où $d_b^c = d_a^c$ et $d_b^t = d_a^t$. Poser $D = D \cup D_b$.
- Si $y > y_z$, alors $P = P \cup c$ et construire $D_c = [d_c^c, d_c^t]$ où $d_c^c = d_a^c$ et $d_c^t = d_a^t$. Poser $D = D \cup D_c$.

4. Poser $N_{A_{y_z}} = P$ et $D_{A_{y_z}} = D$.

La figure 4.4 présente un exemple de l'application de cet algorithme pour un échantillon contenant trois séquences génétiques. Il est intéressant de remarquer que l'étape 4(c)i consistant à ajouter un marqueur à la position π_z sur chaque séquence du graphe G n'est pas exécuté dans cet algorithme. En effet, cette étape est utile à la conceptualisation du problème, mais d'un point de vue strictement computationnel, elle n'est pas nécessaire à l'obtention des trois ensembles d'intérêt.

4.4 Test d'hypothèses

Nous avons présenté dans les sections précédentes la méthodologie permettant d'obtenir une estimation (y_z^*) de la position d'une mutation causale sur une séquence génétique. Cette section a pour but de construire un test d'hypothèses sur notre estimateur, c'est-à-dire de tester s'il existe une association entre le marqueur situé à la position y_z^* , que nous notons $x_{y_z^*}$, et la maladie génétique d'intérêt. Les hypothèses de notre test sont donc :

H₀ : il n'y pas d'association entre $x_{y_z^*}$ et la maladie génétique ;

H₁ : il y a une association entre $x_{y_z^*}$ et la maladie génétique.

Rappelons que la maladie génétique s'exprime dans nos données à travers le vecteur de phénotype Φ , il serait donc plus juste de réécrire les hypothèses précédentes de la façon suivante :

H₀ : il n'y pas d'association entre $x_{y_z^*}$ et $\Phi = \{\phi_1, \dots, \phi_n\}$;

H₁ : il y a une association entre $x_{y_z^*}$ et $\Phi = \{\phi_1, \dots, \phi_n\}$.

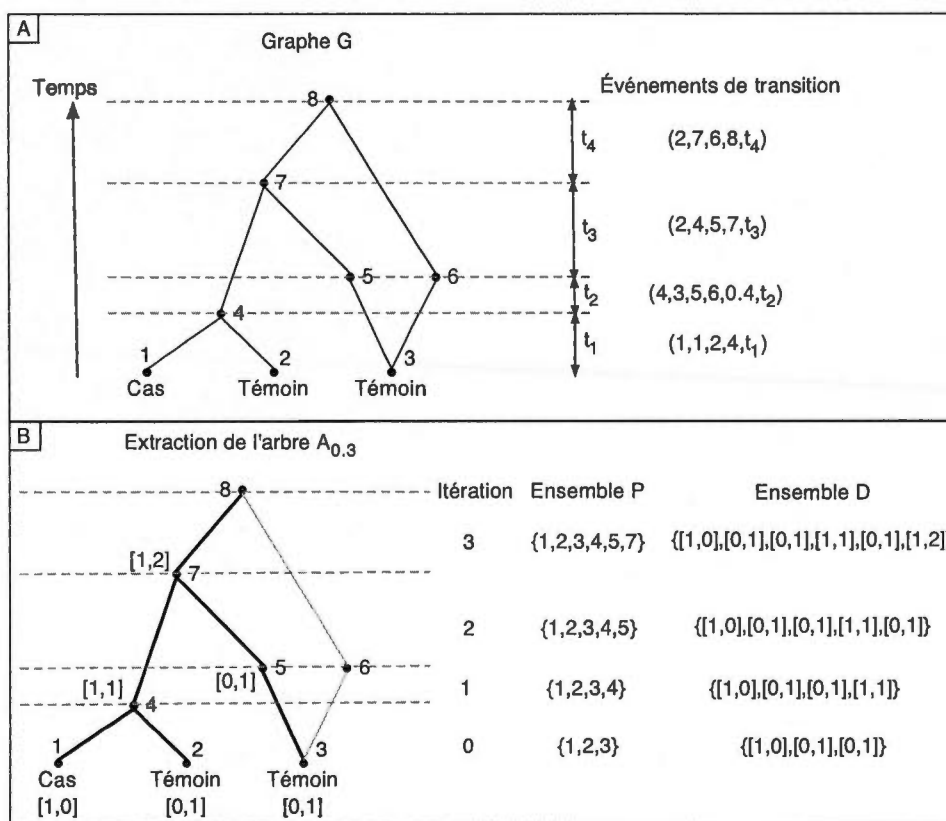


Figure 4.4 Exemple d'application de l'algorithme précédent afin d'obtenir les ensembles $N_{A_{y_z}}$ et $D_{A_{y_z}}$ pour $y_z = 0.3$. Nous retrouvons en A le graphe G où chaque séquence est représentée par un nœud numéroté de 1 à 8. Les 4 événements de transition de ce graphe sont présentés entre parenthèse à la droite du graphe. La notation utilisée est celle illustrée à la figure 4.2. Nous retrouvons en B les 4 itérations permettant d'obtenir les ensembles $N_{A_{0.3}}$ et $D_{A_{0.3}}$. Les branches noires du graphe G représentent l'arbre $A_{0.3}$. La matrice D_x est représentée entre crochet pour chacun des nœuds de l'arbre partiel. Par exemple, les descendants du nœud 7 appartenant à l'échantillon de départ sont les feuilles 1, 2 et 3 ; donc 1 cas et 2 témoins d'où $D_7 = [1, 2]$.

La statistique de test employée pour exécuter ce test d'hypothèses est la vraisemblance $L(c_T = y_z^*; H_0, \Phi)$. Il n'existe malheureusement pas de forme analytique pour la distribution échantillonnale sous H_0 ; il est cependant facile de simuler des valeurs provenant de cette distribution en utilisant un test de permutation (Churchill et Doerge, 1994). En effet, sous H_0 le marqueur $x_{y_z^*}$ est indépendant du vecteur Φ , il suffit donc de calculer $L(c_T = y_z^*; H_0, \Phi_p)$ pour $p = 1, \dots, P$, où P est le nombre de permutations désiré et Φ_p est un vecteur dans lequel les valeurs ϕ_i du vecteur Φ ont été permutées, c'est-à-dire que nous avons permuté l'attribution des cas et des témoins. Suite à l'obtention des valeurs de la vraisemblance sous H_0 , la valeur-p s'obtient simplement en calculant le nombre de valeurs parmi celles-ci qui sont supérieures à $L(c_T = y_z^*; H_0, \Phi)$ et en le divisant par P , ce qui nous donne :

$$\text{valeur} - p = \frac{|\{p | L(c_T = y_z^*; H_0, \Phi_p) > L(c_T = y_z^*; H_0, \Phi)\}|}{P}.$$

Ce test d'hypothèses peut être exécuté très rapidement avec la méthode DMap, car les ARGs étant générés indépendamment du vecteur de phénotype Φ , il est possible d'utiliser un même ARG afin de calculer les valeurs de la vraisemblance pour les P permutations. L'algorithme qui suit est simplement l'algorithme de DMap présenté à la section 4.3.1 auquel nous avons apporté quelques modifications afin d'y introduire les tests de permutation. Ces modifications sont en caractère gras.

1. Choisir un ensemble $Y = (y_1, \dots, y_Z)$ de valeurs possibles et à évaluer pour la position c_T du TIM.
2. Poser H_0 pour l'ensemble contenant les n séquences de notre échantillon ainsi que Φ pour le vecteur contenant le phénotype de chacune des séquences.
3. Construire un ensemble $\Phi' = \{\Phi_1, \Phi_2, \dots, \Phi_P\}$ contenant P vecteurs pour lesquels le phénotype des séquences a été permuté et un vecteur $V = [V_{y_1}, \dots, V_{y_Z}] = [0, \dots, 0]$ qui contiendra les valeur-p des Z tests d'association entre x_{y_z} et la maladie.
4. Pour $i = 1, \dots, M$ faire

(a) Construire un graphe de recombinaison ancestral $G^{(i)}$ à l'aide de la distribution proposée de Fearnhead et Donnelly en suivant les étapes suivantes :

i. Poser $\tau = 0$.

ii. Tant que $|H_\tau^{(i)}| \neq 1$, c'est-à-dire tant que nous n'avons pas atteint le MRCA, faire

- Générer le temps $t_\tau^{(i)}$ avant le prochain événement tel que $t_\tau^{(i)} \sim \exp(\frac{n(n-1+\alpha\theta+\beta\rho)}{2})$, où $n = |H_\tau^{(i)}|$, c'est-à-dire le nombre de séquences présentes à l'état $H_\tau^{(i)}$.
- Calculer $Q(H_{\tau+1}^{(i)}|H_\tau^{(i)})$ à l'aide de l'équation 3.8 pour tous les états $H_{\tau+1}^{(i)}$ consistants avec $H_\tau^{(i)}$, i.e tel que $P(H_\tau^{(i)}|H_{\tau+1}^{(i)}) > 0$.
- Choisir le prochain état $H_{\tau+1}^{(i)}$ de l'ARG proportionnellement à sa pondération.
- Poser $\tau = \tau + 1$.

(b) Calculer la quotient $R^{(i)} = P(G^{(i)})/Q(G^{(i)}) = \prod_{\tau=0}^{\tau^*-1} \pi(H_\tau^{(i)})/\pi(H_{\tau+1}^{(i)})$.

(c) Pour $z = 1, \dots, Z$ faire

- i. Ajouter un marqueur à la position y_z sur chaque séquence du graphe $G^{(i)}$.
- ii. Extraire l'arbre partiel A_{y_z} et construire l'ensemble $B_{A_{y_z}} = \{b_1, \dots, b_{B_z}\}$ contenant les branches de A_{y_z} .
- iii. Pour chaque $b \in B_{A_{y_z}}$ faire
 - Calculer $H_b = P(\Phi|G^{(i)}, y_z, b) \cdot P(b|G^{(i)}, y_z)$ à l'aide des équations 4.6 et 4.8.
 - Pour $p = 1, \dots, P$ calculer $H_b^p = P(\Phi_p|G^{(i)}, y_z, b) \cdot P(b|G^{(i)}, y_z)$ à l'aide des équations 4.6 et 4.8.
- iv. Calculer $P(\Phi|G^{(i)}, y_z) = \sum_{b \in B_{A_{y_z}}} H_b$.
- v. Pour $p = 1, \dots, P$ calculer $P(\Phi_p|G^{(i)}, y_z) = \sum_{b \in B_{A_{y_z}}} H_b^p$.

5. Pour $z = 1, \dots, Z$

- Calculer $L(c_T = y_z; H_0, \Phi) = \frac{1}{M} \sum_{i=1}^M R^{(i)} \cdot P(\Phi|G^{(i)}, y_z)$.

- Pour $p = 1, \dots, P$ faire
 - Calculer $L(c_T = y_z; H_0, \Phi_p) = \frac{1}{M} \sum_{i=1}^M R^{(i)} \cdot P(\Phi_p | G^{(i)}, y_z)$.
 - Si $L(c_T = y_z; H_0, \Phi_p) > L(c_T = y_z; H_0, \Phi)$, poser $V_{y_z} = V_{y_z} + 1$.
- 6. L'estimateur de la position c_T est $y_z^* \in Y$ tel que $\max\{L(y_1), \dots, L(y_Z)\} = L(y_z^*)$.
- 7. La valeur-p du test d'association entre $x_{\pi_z^*}$ et la maladie est $V_{\pi_z^*}/P$.

Un exemple de résultat que nous pouvons obtenir grâce à l'algorithme précédent est illustré à la figure 4.5. Le graphique de gauche représente le logarithme de la fonction de vraisemblance $L(c_T)$ tandis que le graphique de droite représente la valeur-p du test d'association entre $x_{y_z^*}$ et la maladie (point vert); à des fins d'illustration, nous avons aussi représenté les valeurs-p des tests d'association entre $x_{y_i}, y_i \in Y$ et la maladie.

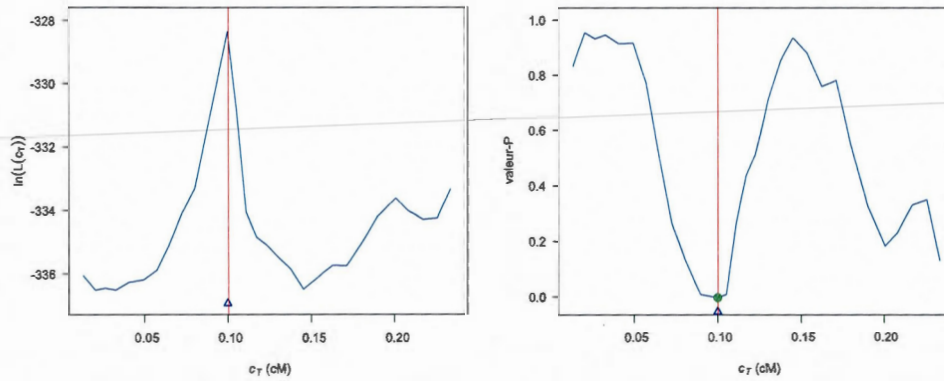


Figure 4.5 Graphiques obtenus à l'aide de la méthode DMap. Le graphique de droite illustre le logarithme de la fonction de vraisemblance évaluée aux positions $y_i \in Y$, tandis que le graphique de gauche illustre les valeurs-p obtenues lors du test d'association entre chacune de ses valeurs et la maladie. Le point vert représente la valeur-p du test d'association entre la position estimée par DMap (y_z^*) et la maladie. La ligne verticale rouge indique la vraie position du TIM tandis que le triangle bleu indique la position y_z^* correspondant à l'estimation par maximum de vraisemblance obtenu avec la méthode DMap. Pour cet exemple, la position estimée est très près de la vraie position et elle correspond de plus à la valeur ayant la plus petite valeur-p.

CHAPITRE V

RÉSULTATS

Dans les deux derniers chapitres de ce mémoire, nous avons présenté des méthodes de cartographie génétique fine déjà existantes (MapARG, Margarita et LATAG) ainsi que la nouvelle méthode proposée dans ce mémoire (DMap). Toutes ces méthodes ont le même objectif : estimer l'emplacement sur une séquence génétique d'une mutation influençant une maladie génétique. Nous sommes donc rendus à l'étape où il serait intéressant de pouvoir analyser les résultats obtenus avec la méthode DMap ainsi que de les comparer avec ceux obtenus par d'autres méthodes similaires. Pour ce faire, nous allons travailler avec des données que nous avons nous-même simulées et qui répondent aux suppositions de notre modèle ; cette pratique est couramment employée en statistique génétique, par exemple Minichiello et Durbin (2006) et Zöllner et Pritchard (2005) utilisent le même procédé. Un avantage considérable à travailler avec des données simulées est que l'on connaît la vraie position de la mutation causale dans nos échantillons, il est donc facile de vérifier si nos estimations sont bonnes.

Dans ce chapitre, nous allons premièrement décrire la façon avec laquelle nous simulons nos données, nous analyserons par la suite les performances de la méthode DMap et nous comparerons celles-ci aux performances de la méthode MapARG, de la méthode Margarita et d'une méthode basée sur les études d'association qui sera décrite brièvement dans ce chapitre. Nous terminerons ce chapitre avec une discussion sur les résultats obtenus ainsi que sur les améliorations possibles et les possibilités de développement de la méthode DMap.

5.1 Simulation des données

Tel que mentionné précédemment, nous allons utiliser des données simulées afin de tester l'efficacité de la méthode DMap. Il y a deux étapes principales afin d'obtenir un échantillon de séquences génétiques. Il faut premièrement simuler une population de séquences génétiques ; ceci se fait facilement grâce au programme *ms* (Hudson, 2002). Ce programme permet de générer des graphes de recombinaison ancestraux selon la théorie de la coalescence avec recombinaison présentée au deuxième chapitre. Rappelons qu'une fois un ARG généré, nous obtenons un ensemble de séquences génétiques qui peuvent être représentées sous forme de SNP grâce au modèle des sites infinis pour les mutations (voir les figures 2.4(a) et 2.4(b)). De plus, il est possible de spécifier au programme *ms* la taille désirée de la population (supposée fixe au cours des générations), le nombre de marqueurs constituant chaque séquence ainsi que la longueur des séquences (r) voulus. La deuxième étape consiste à créer un échantillon de cas et de témoins à partir d'une population de séquences génétiques. ~~Pour ce faire, nous avons utilisé un programme~~ C++ appelé *Sample*, auquel nous avons fourni la fréquence de la mutation causale dans la population (μ_p), la fonction de pénétrance $F = (f_0, f_1)$ ainsi que le nombre de cas (n_c) et de témoins (n_t) voulu. Les principales étapes de l'application du programme *Sample* sont les suivantes :

- sélectionner un marqueur dont l'allèle mutant ($\ll 1 \gg$) a une fréquence près de la valeur de μ_p : ce marqueur sera le TIM ;
- assigner un statut de cas ou de témoin à chaque séquence selon son allèle (muté ou non) au marqueur du TIM et la fonction de pénétrance F ;
- retirer le marqueur représentant le TIM pour chaque séquence et garder son emplacement en mémoire : on connaît donc la vraie position du TIM ;
- éliminer les marqueurs possédant un allèle dont la fréquence est inférieure à 1% : de tels marqueurs ne sont pas informatifs, car plus de 99% de la population ont les mêmes allèles à ces marqueurs ;
- choisir aléatoirement n_c séquences auxquelles nous avons assigné un statut de cas et n_t séquences auxquelles nous avons assigné un statut de témoin.

Nous venons de décrire brièvement la procédure à suivre afin d'obtenir un échantillon de données à partir d'une population de séquences génétiques ; notons que cet échantillon n'est pas aléatoire, puisque le nombre de cas et de témoins désiré est fixé à l'avance.

Afin d'analyser les résultats obtenus avec la méthode DMap, nous avons simulé 100 populations de 10 000 séquences génétiques de longueur $r = 0.25 \text{ Mb} \approx 0.25 \text{ cM}$ et contenant 1 000 marqueurs. Par la suite, pour chacune de ces populations, nous avons généré 4 échantillons selon 4 scénarios d'intérêt. Nous avons par la suite généré 100 autres populations de 10 000 séquences génétiques et contenant 1 000 marqueurs, mais cette fois-ci contenant des séquences de longueur $r = 0.625 \text{ Mb} \approx 0.625 \text{ cM}$. Encore une fois, pour chacune de ces 100 populations, nous avons généré 4 échantillons selon 4 autres scénarios d'intérêt. Le tableau 5.1 présente les différents scénarios que nous avons considéré afin de générer nos échantillons. Notons que RR , qui signifie risque relatif, représente le risque qu'une séquence provienne d'un individu atteint par la maladie d'intérêt relativement au fait qu'elle soit porteuse ou non du TIM. Ce risque relatif est simplement le ratio de la probabilité qu'une séquence porteuse du TIM provienne d'un cas sur la probabilité qu'une séquence non porteuse du TIM provienne d'un cas, donc $RR = f_1/f_0$. Ainsi, plus la valeur de RR est élevée, plus la mutation a une grande influence sur la maladie ; nous pensons donc que les méthodes de cartographie génétique devraient donner de meilleurs résultats lorsque le risque relatif est élevé.

5.2 Résultats obtenus avec la méthode DMap

5.2.1 Les paramètres de simulation

Mentionnons tout d'abord qu'à moins d'avis contraire, tous les résultats obtenus par la méthode DMap qui sont présentés dans ce chapitre ont été produit en utilisant les paramètres décrits dans cette sous-section. Les séquences de nos échantillons contiennent plus d'une centaine de marqueurs et il n'est pas possible de construire des ARGs en un temps raisonnable pour des séquences contenant autant de marqueurs, nous avons donc choisi d'utiliser $L = 64$ marqueurs par séquence pour les échantillons provenant

Population 1 à 100 : $r = 0.25$ cM						
Scénario	t_m	$F = (f_0, f_1)$	RR	n_c	n_t	n
A	0.0189	$F = (0.01, 0.099)$	9.9	200	200	400
B	0.108	$F = (0.059, 0.549)$	9.3	200	200	400
C	0.0222	$F = (0.016, 0.078)$	4.875	200	200	400
D	0.222	$F = (0.21, 0.33)$	1.5	200	200	400
Population 101 à 200 : $r = 0.625$ cM						
Scénario	t_m	$F = (f_0, f_1)$	RR	n_c	n_t	n
F	0.0705	$F = (0.06, 0.165)$	2.75	200	200	400
G	0.0842	$F = (0.082, 0.104)$	1.27	200	200	400
H	0.012	$F = (0.011, 0.0315)$	2.86	200	200	400
I	0.011	$F = (0.0105, 0.021)$	2	200	200	400

Tableau 5.1 Tableau résumant les caractéristiques des échantillons générés pour chacune des 200 populations simulées. La fréquence de la maladie dans la population est représentée par t_m , n correspond au nombre de séquences dans l'échantillon, n_t au nombre de séquences provenant d'individus non affectés par la maladie et n_c au nombre de séquences provenant d'individus affectés par la maladie.

des populations 1 à 100 et $L = 50$ marqueurs pour les échantillons provenant des populations 101 à 200. La façon dont nous choisissons ces marqueurs est très simple : pour chaque échantillon considéré, nous sélectionnons tout d'abord, parmi tous les marqueurs présents sur les séquences de l'échantillon, les 128 marqueurs les plus polymorphiques, c'est à dire les marqueurs dont la fréquence de chacun des deux allèles est près de 50%. Par la suite, nous sélectionnons les L marqueurs les plus équidistants parmi ces 128 marqueurs. Cette procédure nous assure de sélectionner des marqueurs informatifs et bien répartis sur nos séquences génétiques. De plus, l'ensemble $\pi = \{\pi_1, \dots, \pi_Z\}$, contenant les Z positions dont on veut calculer la vraisemblance, est formé de sorte à ce qu'il contienne $Z = L - 1$ valeurs correspondant aux milieux des $L - 1$ intervalles formés par les L marqueurs constituant nos séquences. Finalement, le nombre de graphes de

recombinaison ancestraux M utilisé dans le calcul de la fonction de vraisemblance est de 50 ; ce nombre peut sembler petit a priori, mais plusieurs tests (non présentés dans ce mémoire) nous permettent d'affirmer qu'il est suffisamment grand pour obtenir une bonne approximation de la fonction de vraisemblance.

5.2.2 Comparaison des différentes versions de DMap

Nous allons tout d'abord comparer les résultats obtenus avec les différentes versions de la fonction de vraisemblance que nous avons présentées à la section 4.2. Rappelons brièvement quelles étaient ces versions. Pour cela, il est utile de réécrire l'équation 4.3 représentant la forme classique de la vraisemblance :

$$L(c_T) \approx \frac{1}{M} \sum_{i=1}^M \frac{P(\Phi|G^{(i)}, c_T) \cdot P(G^{(i)})}{Q(G^{(i)})}.$$

Nous avons montré à la section 4.2.1 que le poids de chaque graphe ($P(G^{(i)})/Q(G^{(i)})$) se calculait à l'aide de l'équation suivante :

$$\frac{P(G^{(i)})}{Q(G^{(i)})} = \prod_{\tau=0}^{\tau^*-1} \left[\frac{\pi(H_\tau^{(i)})}{\pi(H_{\tau+1}^{(i)})} \right], \quad (5.1)$$

ou qu'il pouvait être considéré comme une constante. De plus, à la section 4.2.2, nous avons présenté une équation exacte (4.8) et une équation approximative (4.9) pour la probabilité $P(b|G, c_T)$. Les quatre versions possibles pour la vraisemblance sont donc :

$$L(c_T) \approx \frac{1}{M} \sum_{i=1}^M \left(\sum_{b \in B_{A_{c_T}}} \Psi_b \cdot \frac{P(G^{(i)})}{Q(G^{(i)})} \right), \quad (5.2)$$

$$L(c_T) \propto \frac{1}{M} \sum_{i=1}^M \left(\sum_{b \in B_{A_{c_T}}} \Psi_b \right), \quad (5.3)$$

$$L(c_T) \propto \frac{1}{M} \sum_{i=1}^M \left(\sum_{b \in B_{A_{c_T}}} \Psi_b \cdot P(N_m = n_m^{(b)}) \right), \quad (5.4)$$

$$L(c_T) \approx \frac{1}{M} \sum_{i=1}^M \left(\sum_{b \in B_{A_{c_T}}} \Psi_b \cdot P(N_m = n_m^{(b)}) \cdot \frac{P(G^{(i)})}{Q(G^{(i)})} \right), \quad (5.5)$$

où $\Psi_b = P(\Phi|G, c_T, b) \frac{|b|}{\sum_{w \in B_{A_{c_T}}} |w|}$. Les équations 5.2 et 5.3 représentent des versions de la fonction de vraisemblance dans lesquelles la forme exacte de la probabilité de

poser une mutation sur une branche b est employée, tandis que les équations 5.4 et 5.5 contiennent la forme approximative de cette probabilité. Dans les équations 5.2 et 5.5, les poids des graphes sont calculés à l'aide de l'équation 5.1 tandis que dans les équations 5.3 et 5.4, ils sont considérés comme équivalents.

Afin de comparer les résultats obtenus avec ces quatre versions de la fonction de vraisemblance, nous présentons dans les pages qui suivent (voir figures 5.1 à 5.4) les graphiques illustrant les fonctions de vraisemblance obtenus avec chacune des versions pour les échantillons, générés selon les scénarios A, B, C et D, de la population 54. Nous pouvons voir sur chacun des différents graphiques le logarithme de la fonction de vraisemblance (ligne bleue), le vrai emplacement du TIM (ligne verticale rouge) ainsi que l'estimation obtenue par la méthode DMap (triangle bleu). En analysant les figures 5.1, 5.2, 5.3 et 5.4, nous remarquons immédiatement que les deux graphiques situés au haut de chaque figure, qui correspondent aux résultats obtenus avec les équations 5.2 et 5.3, illustrent une mauvaise performance de la méthode DMap, mis à part la figure 5.2. Quant à eux, les graphiques du bas de chaque figure, qui correspondent aux résultats obtenus avec les équations 5.4 et 5.5, semblent montrer de bien meilleurs résultats. En effet, pour la plupart de ceux-ci, l'estimation faite par DMap semble très près de la vraie position du TIM (le triangle bleu est positionné près de la ligne verticale rouge). Mentionnons que nous obtenons des résultats similaires avec les 99 autres populations, nous pouvons donc aisément conclure que la modification que nous avons apportée à la probabilité $P(b|G, c_T)$ est très utile; pour la suite de ce mémoire, nous adopterons l'équation approximative 4.9 afin de calculer la probabilité $P(b|G, c_T)$. Nous devons maintenant vérifier si la méthode DMap produit de meilleur résultat lorsque l'on utilise l'équation 5.1 afin de calculer les poids des graphes dans notre fonction de vraisemblance plutôt que de les considérer équivalents. Pour ce faire, nous avons choisi de comparer la distribution du biais de notre estimateur obtenue en utilisant la version de la fonction de vraisemblance de l'équation 5.4 à celle obtenue en utilisant l'équation 5.5. Nous avons donc calculé ces biais pour les scénarios A, B, C et D des populations 1 à 100. Les différentes distributions obtenues selon la version de l'estimateur et le scénario (A, B, C ou D) sont

illustrées par des diagrammes en boîtes à la figure 5.5. En observant les 4 graphiques de cette figure, nous pouvons constater que les distributions des biais obtenus par les équations 5.4 et 5.5 semblent centrés en 0, comme nous l'espérons. Cependant, les distributions produites par l'équation 5.4 semblent légèrement avoir une moins grande variabilité. La méthode DMap semble donc plus performante lorsque nous considérons que tous les graphes générés ont le même poids dans notre équation de vraisemblance. Pour la suite de ce mémoire, nous allons donc supposer que les graphes générés sont équiprobables et utiliser la fonction de vraisemblance de l'équation 5.4.

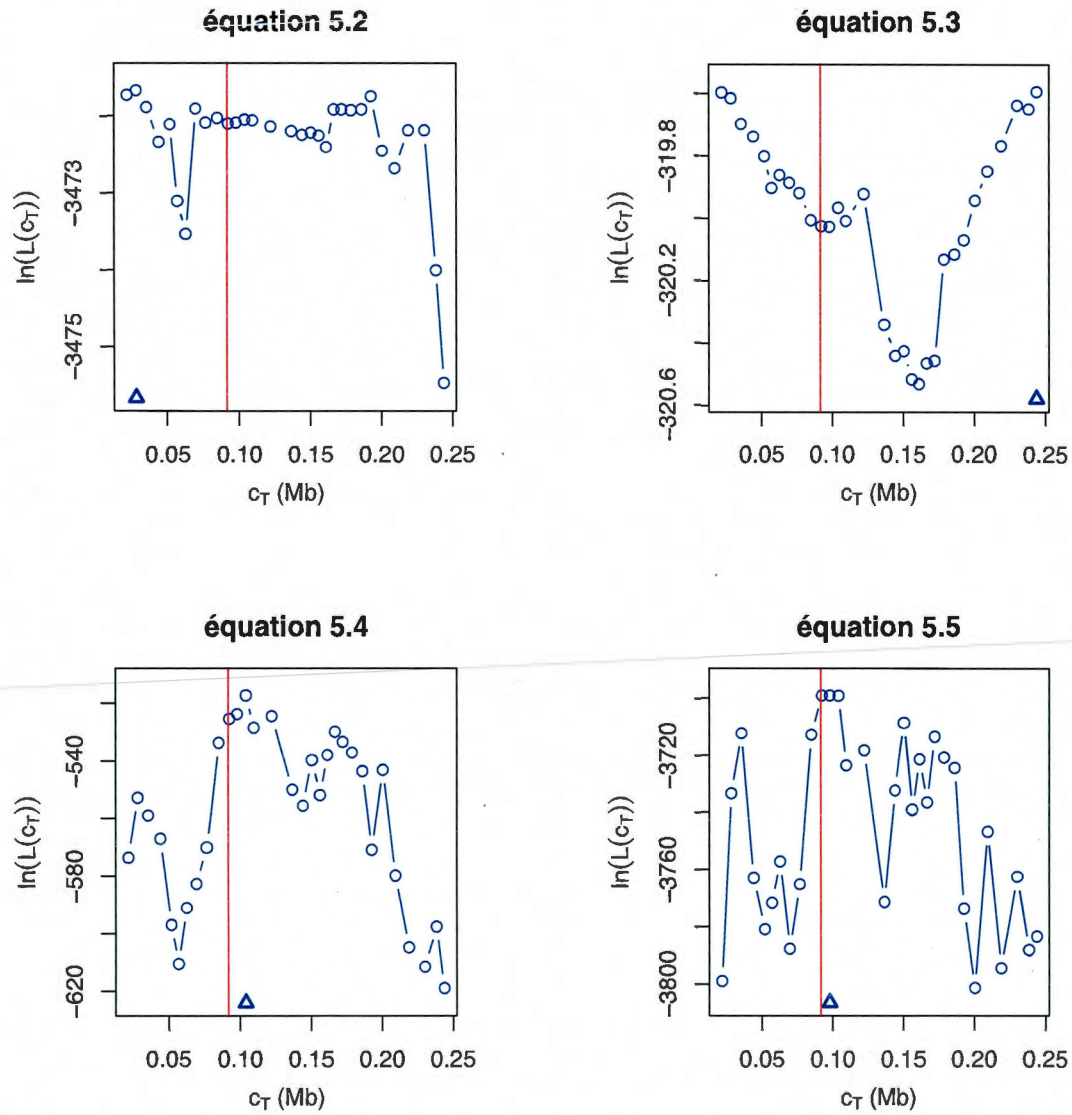


Figure 5.1 Figure illustrant les graphiques obtenus pour chacune des quatre versions de la fonction de vraisemblance. L'échantillon généré selon le scénario A de la population 54 a été utilisé afin d'obtenir ces graphiques.

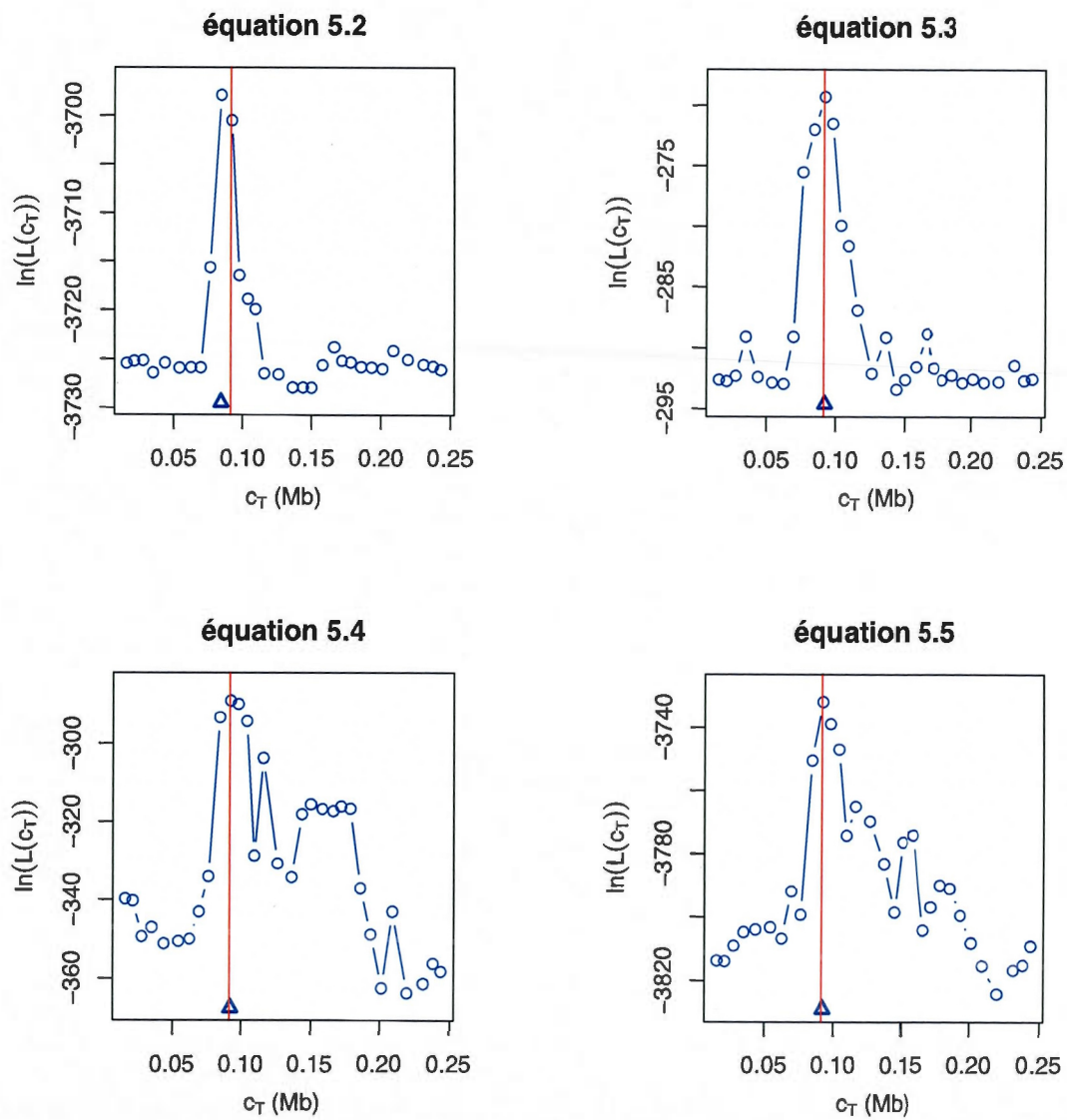


Figure 5.2 Figure illustrant les graphiques obtenus pour chacune des quatre versions de la fonction de vraisemblance. L'échantillon généré selon le scénario B de la population 54 a été utilisé afin d'obtenir ces graphiques.

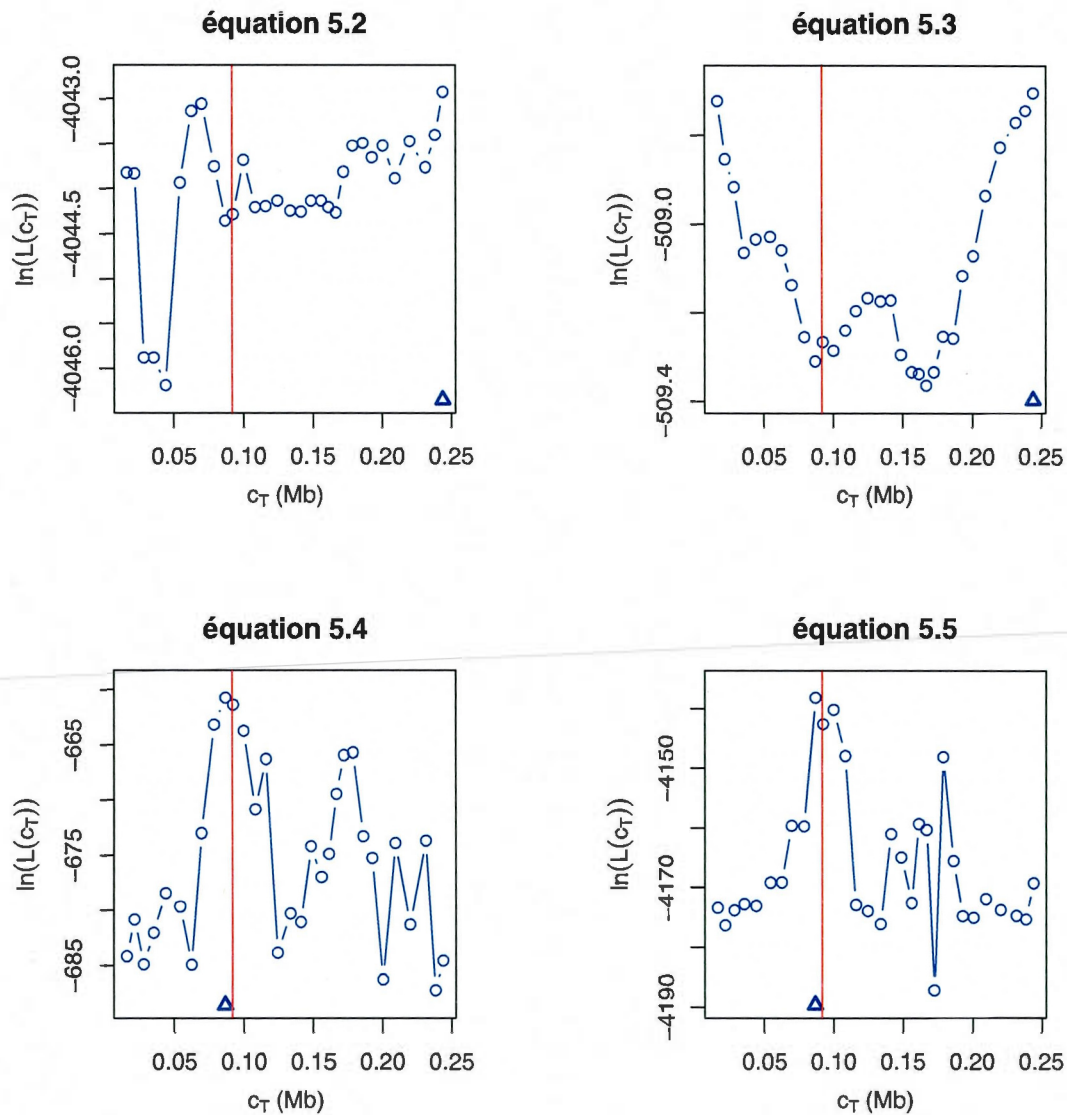


Figure 5.3 Figure illustrant les graphiques obtenus pour chacune des quatre versions de la fonction de vraisemblance. L'échantillon généré selon le scénario C de la population 54 a été utilisé afin d'obtenir ces graphiques.

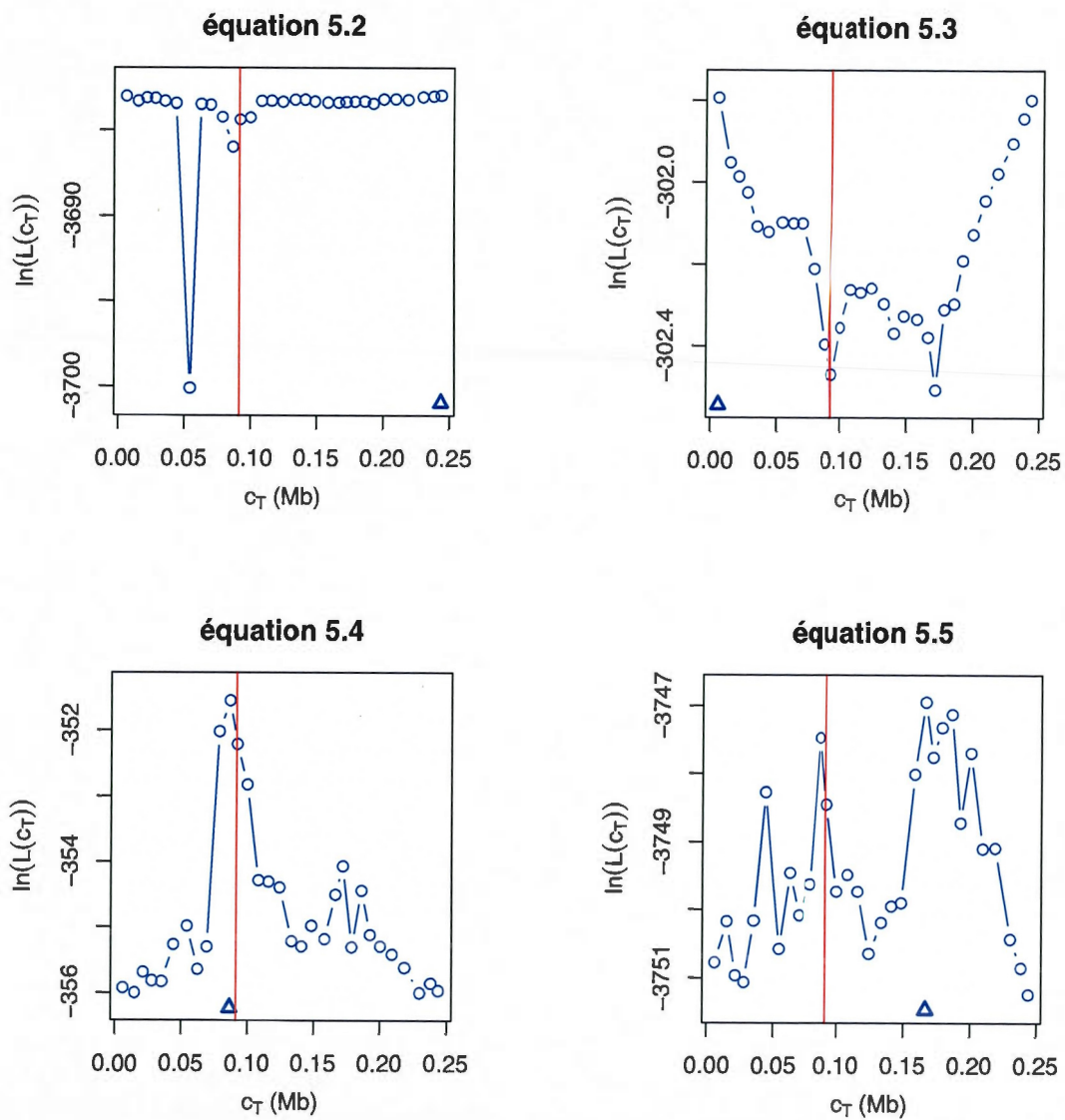


Figure 5.4 Figure illustrant les graphiques obtenus pour chacune des quatre versions de la fonction de vraisemblance. L'échantillon généré selon le scénario D de la population 54 a été utilisé afin d'obtenir ces graphiques.

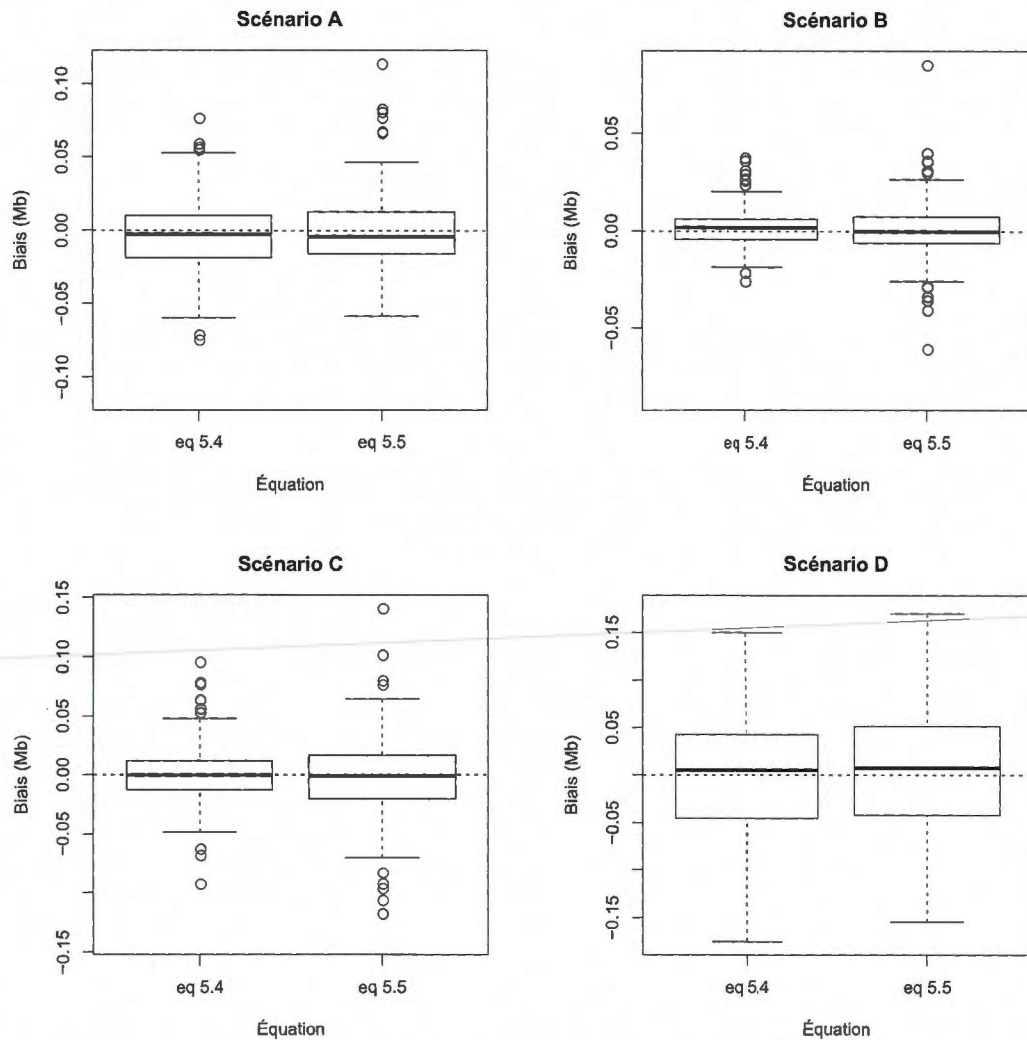


Figure 5.5 Figure illustrant les distributions des biais des estimateurs obtenus avec les deux versions de la fonction de vraisemblance à comparer. Pour chaque scénario (A, B, C et D), la distribution correspondant à l'équation 5.4 semble moins étendue que celle correspondant à l'équation 5.5.

5.2.3 Tests d'hypothèses

Afin de tester la significativité des estimations obtenues avec l'équation 5.4 sur les 4 échantillons de la population 54, nous avons utilisé le test d'hypothèses présenté à la section 4.4. Les valeur-p de nos tests ont été calculées à l'aide de $P = 1\,000$ permutations. Le tableau 5.2 présente les valeur-p obtenues pour les 4 tests d'hypothèses effectués. En observant les valeurs qu'il contient, nous pouvons, pour chacun des tests, rejeter à un seuil $\alpha = 0.05$ l'hypothèse nulle stipulant qu'il n'y a pas d'association entre le marqueur situé à la position de notre estimation pour l'emplacement du TIM et la maladie génétique considérée.

Population 54	
Échantillons	valeur-p
A	< 0.001
B	< 0.001
C	< 0.001
D	0.006

Tableau 5.2 Tableau contenant les valeur-p obtenues suite au test d'hypothèse sur les 4 échantillons de la population 54.

5.3 Comparaison de DMap avec trois autres méthodes de cartographie génétique

Dans cette section, nous allons comparer les performances de la méthode DMap à celles de la méthode MapARG (section 3.2) et de la méthode Margarita (section 3.3). Nous ne comparerons malheureusement pas notre méthode à celle de LATAG (section 3.4), car nous n'avons pas été en mesure d'utiliser le programme de cette méthode. Par contre, nous allons aussi comparer nos résultats à ceux obtenus par une méthode classique en cartographie génétique qui est basée sur l'étude d'association.

5.3.1 Méthode basée sur l'étude d'association

Avant de présenter les résultats obtenus avec les différentes méthodes, nous allons expliquer brièvement la méthode basée sur l'étude d'association. Le but d'une étude d'association est, tout comme le but d'une méthode de cartographie génétique fine, d'estimer l'emplacement d'une mutation causale sur une séquence génétique. Une telle étude consiste à tester l'association entre un marqueur d'une séquence génétique et le phénotype lié à la maladie génétique étudiée, et ce, pour tous les marqueurs contenus sur les séquences génétiques d'un échantillon. Une forte association entre un marqueur et une maladie génétique indique que le marqueur a une grande probabilité d'être situé tout près du marqueur sur lequel le TIM se situe. Cette dernière affirmation est basée sur le *déséquilibre de liaison* ; on dit qu'il y a présence de déséquilibre de liaison dans une population de séquences génétiques lorsqu'il y a une structure non-aléatoire des combinaisons des différents allèles de deux marqueurs d'une séquence génétique. Tel que ~~mentionné au premier chapitre~~, lorsque deux marqueurs sont situés très près l'un de l'autre sur une séquence génétique, il est peu probable qu'il y ait eu une recombinaison entre ces deux marqueurs au cours de l'histoire, il est donc fort probable qu'il y ait un déséquilibre de liaison entre ces deux marqueurs. Ainsi, il nous est possible d'affirmer que les marqueurs situés tout près du TIM devraient eux aussi démontrer une forte association avec la maladie génétique d'intérêt.

Il existe plusieurs méthodes basées sur les études d'association, celle que nous avons choisi d'utiliser dans ce chapitre est celle employant des statistiques du χ^2 afin de calculer l'association entre un marqueur et le phénotype. Cette méthode est très simple et rapide. En effet, connaissant pour chaque séquence d'un échantillon son phénotype ainsi que son allèle pour un marqueur donné, nous obtenons facilement pour chaque marqueur présent sur les séquences de notre échantillon une table de contingence ayant la forme du tableau 5.3.

Une fois que la statistique du χ^2 a été calculée pour tous les marqueurs de nos séquences génétiques, il ne reste plus qu'à calculer la valeur-p associée à chacune de ces statistiques.

L'estimation de la position du TIM par cette méthode est simplement la position du marqueur ayant la plus petite valeur-p.

Allèle / Phénotype	Cas	Témoin	Totaux
Mutant («1»)	x	y	$x + y$
Non-mutant («0»)	w	z	$w + z$
Totaux	200	200	400

Tableau 5.3 Table de contingence utilisée pour des tests d'association entre une maladie génétique et un marqueur donné.

5.3.2 Illustration des résultats obtenus avec chaque méthode

Dans cette sous-section, nous allons illustrer quelques résultats obtenus avec les méthodes DMap, MapARG, Margarita et la méthode présentée dans la sous-section précédente, que nous appellerons Chi2 par souci de simplicité. Nous avons choisi d'illustrer les résultats obtenus avec les scénarios A à D de la population 28 (figures 5.6 à 5.9) et avec les scénarios F à I de la population 105 (figures 5.10 à 5.13), car ceux-ci sont représentatifs du comportement des différentes méthodes. Sur chacune des figures, le graphique en haut à gauche est obtenu avec la méthode DMap, celui en haut à droite est obtenu avec la méthode Margarita, celui en bas à gauche est obtenu avec la méthode MapARG et finalement celui en bas à droite est obtenu avec la méthode Chi2. Rappelons que l'axe des x représente des positions le long d'une séquence génétique. De plus, l'axe des y représente le logarithme de la fonction de vraisemblance pour les méthodes DMap et MapARG tandis qu'il représente moins dix fois le logarithme de la valeur-p pour les méthodes Margarita et Chi2.

En observant les résultats obtenus pour les scénarios A à D (figures 5.6 à 5.9), nous pouvons constater que DMap semble donner de bons et de meilleurs résultats que les 3 autres méthodes. Par contre, lorsque nous observons les résultats obtenus pour les scénarios F à I (figures 5.10 à 5.13), les performances de DMap semblent être quelque peu moins bonnes. Afin de généraliser les résultats illustrés sur ces figures, la sous-section

suivante contient une analyse détaillée des résultats obtenus avec les 4 méthodes.

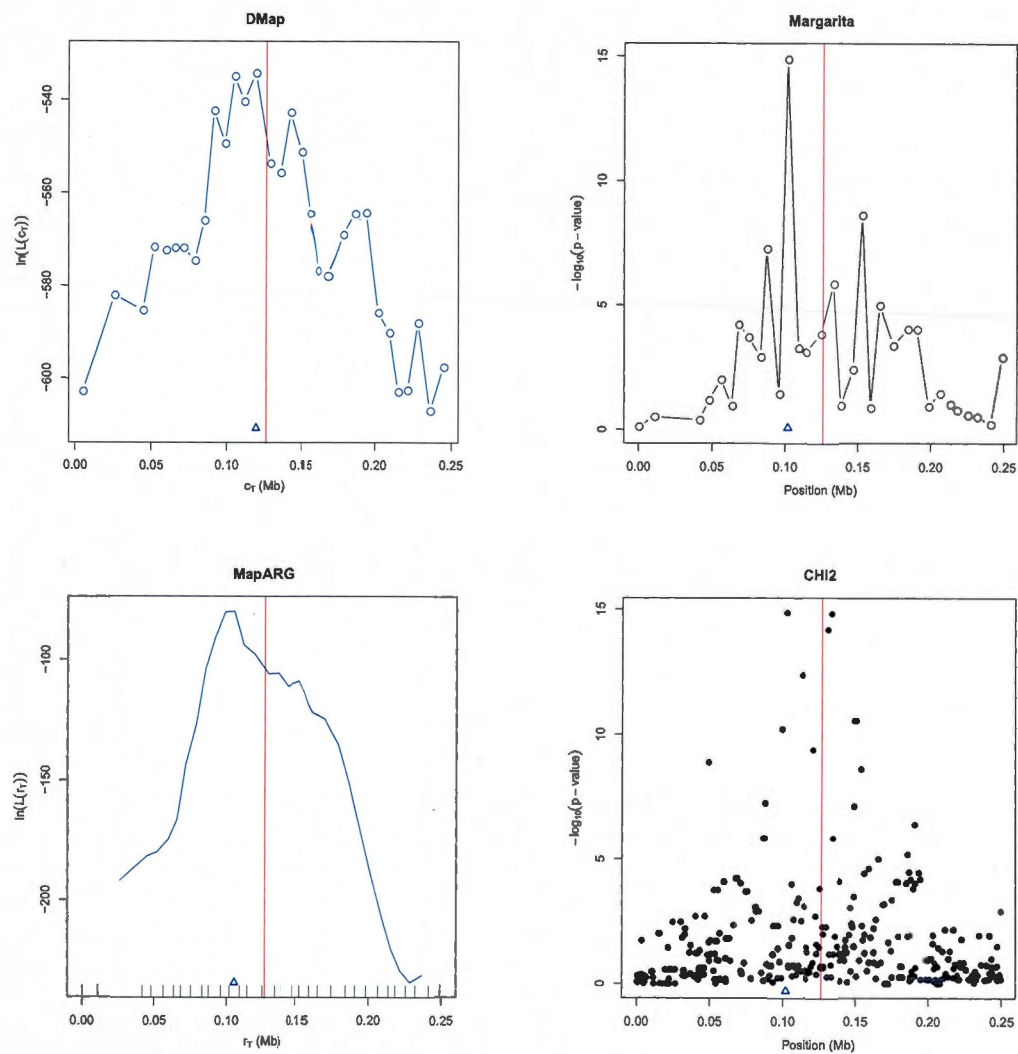


Figure 5.6 Figure illustrant les résultats obtenus avec les 4 méthodes de cartographie à comparer. L'échantillon généré selon le scénario A de la population 58 a été utilisé afin d'obtenir ces résultats.

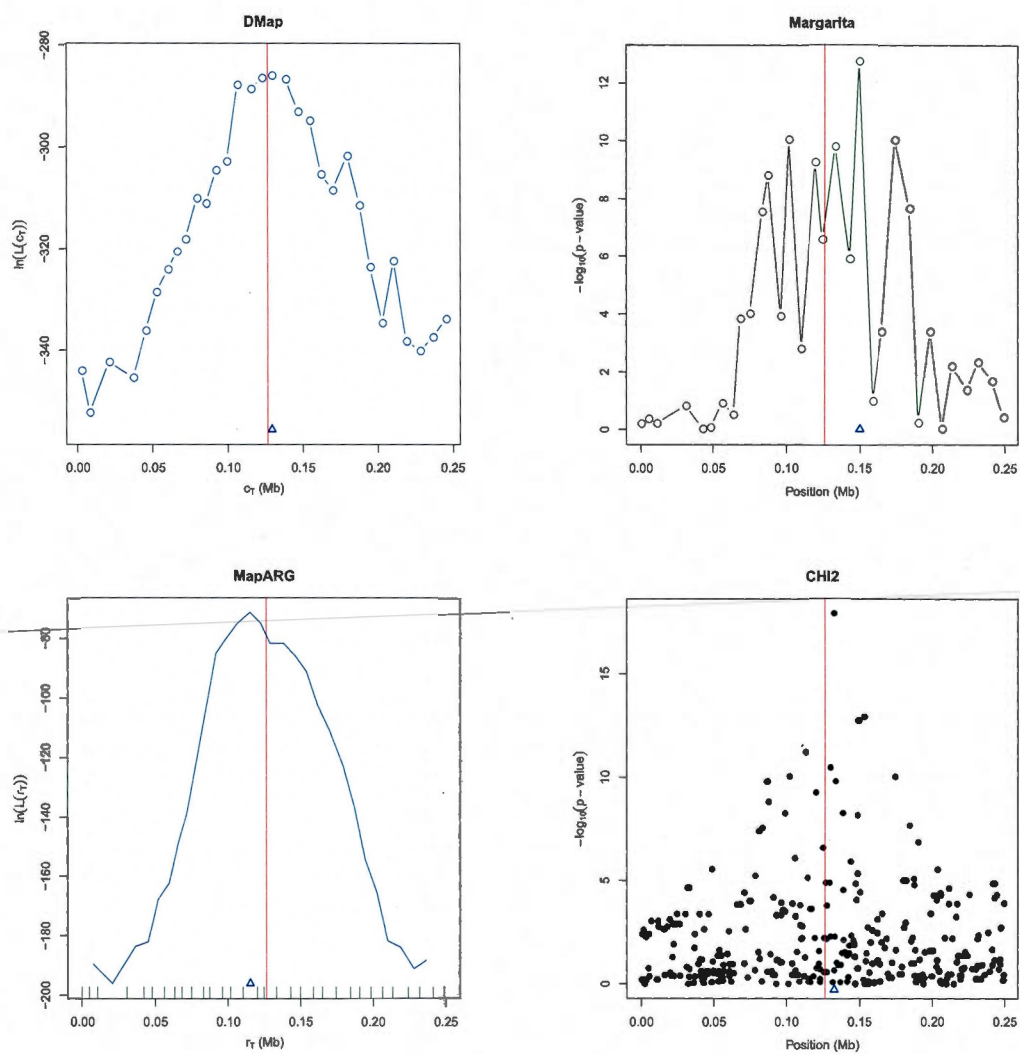


Figure 5.7 Figure illustrant les résultats obtenus avec les 4 méthodes de cartographie à comparer. L'échantillon généré selon le scénario B de la population 58 a été utilisé afin d'obtenir ces résultats.

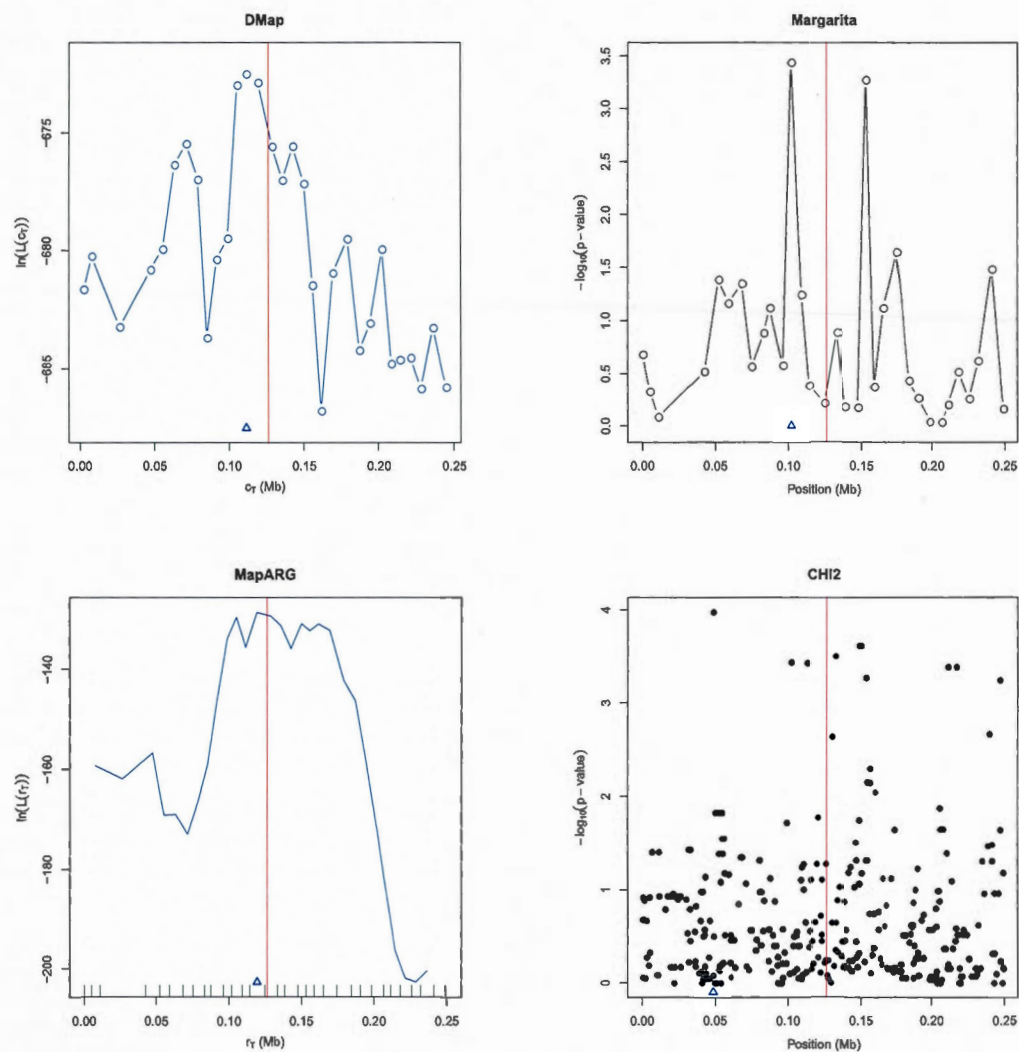


Figure 5.8 Figure illustrant les résultats obtenus avec les 4 méthodes de cartographie à comparer. L'échantillon généré selon le scénario C de la population 58 a été utilisé afin d'obtenir ces résultats.

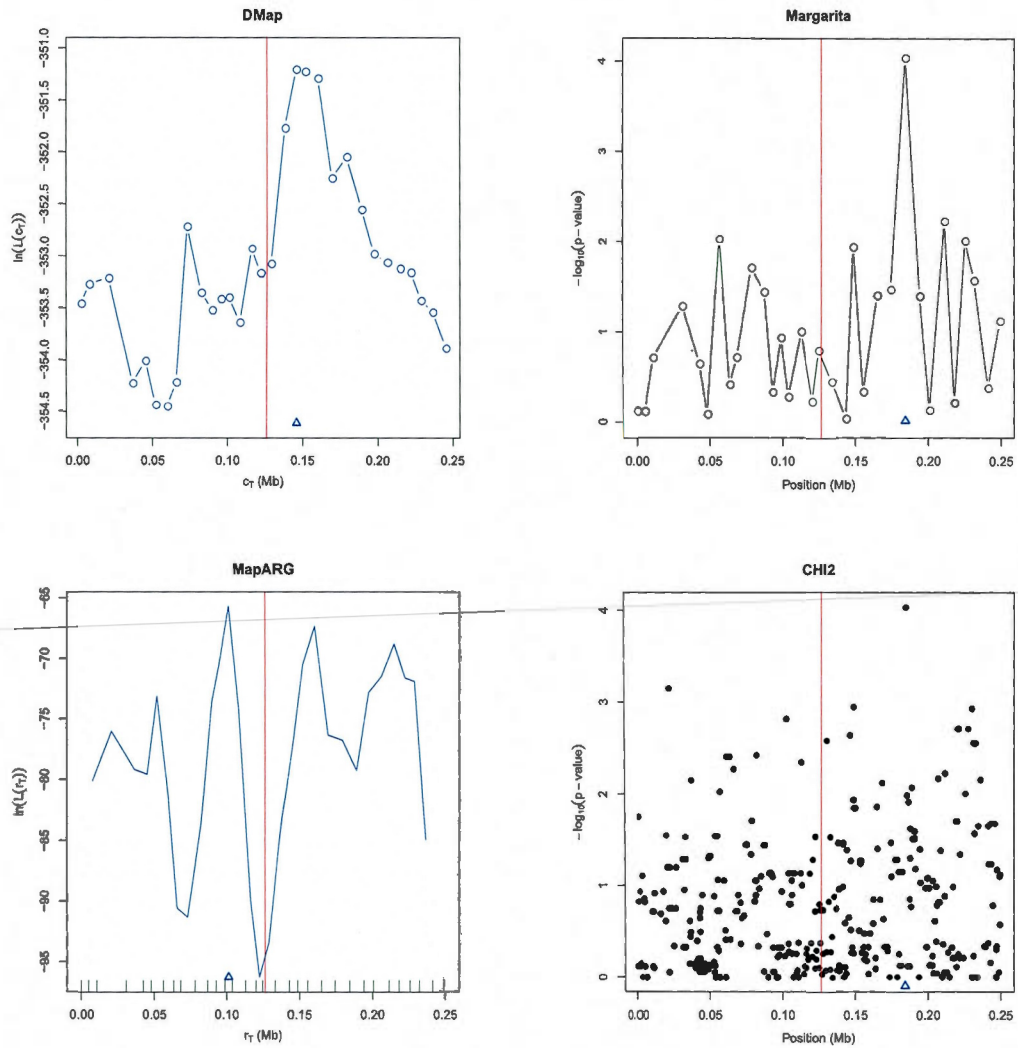


Figure 5.9 Figure illustrant les résultats obtenus avec les 4 méthodes de cartographie à comparer. L'échantillon généré selon le scénario D de la population 58 a été utilisé afin d'obtenir ces résultats.

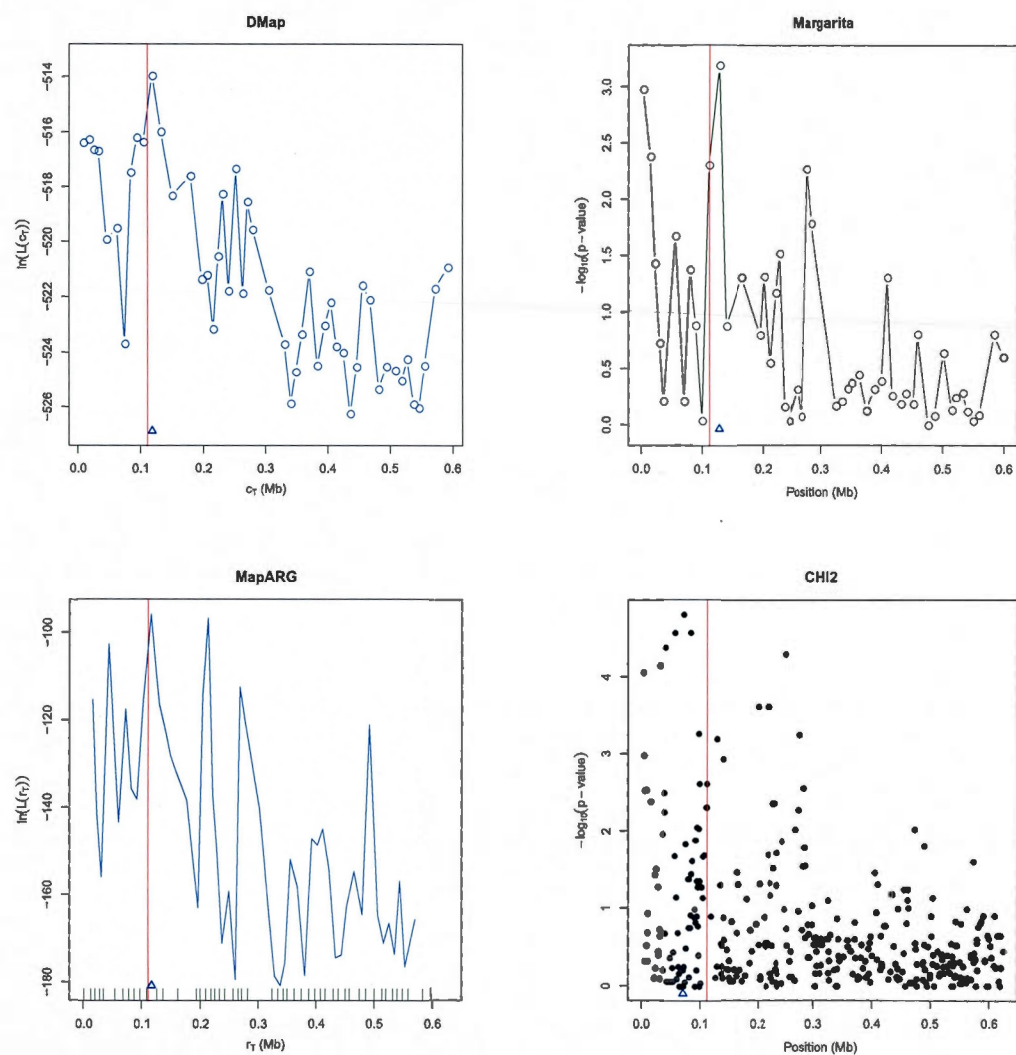


Figure 5.10 Figure illustrant les résultats obtenus avec les 4 méthodes de cartographie à comparer. L'échantillon généré selon le scénario F de la population 105 a été utilisé afin d'obtenir ces résultats.

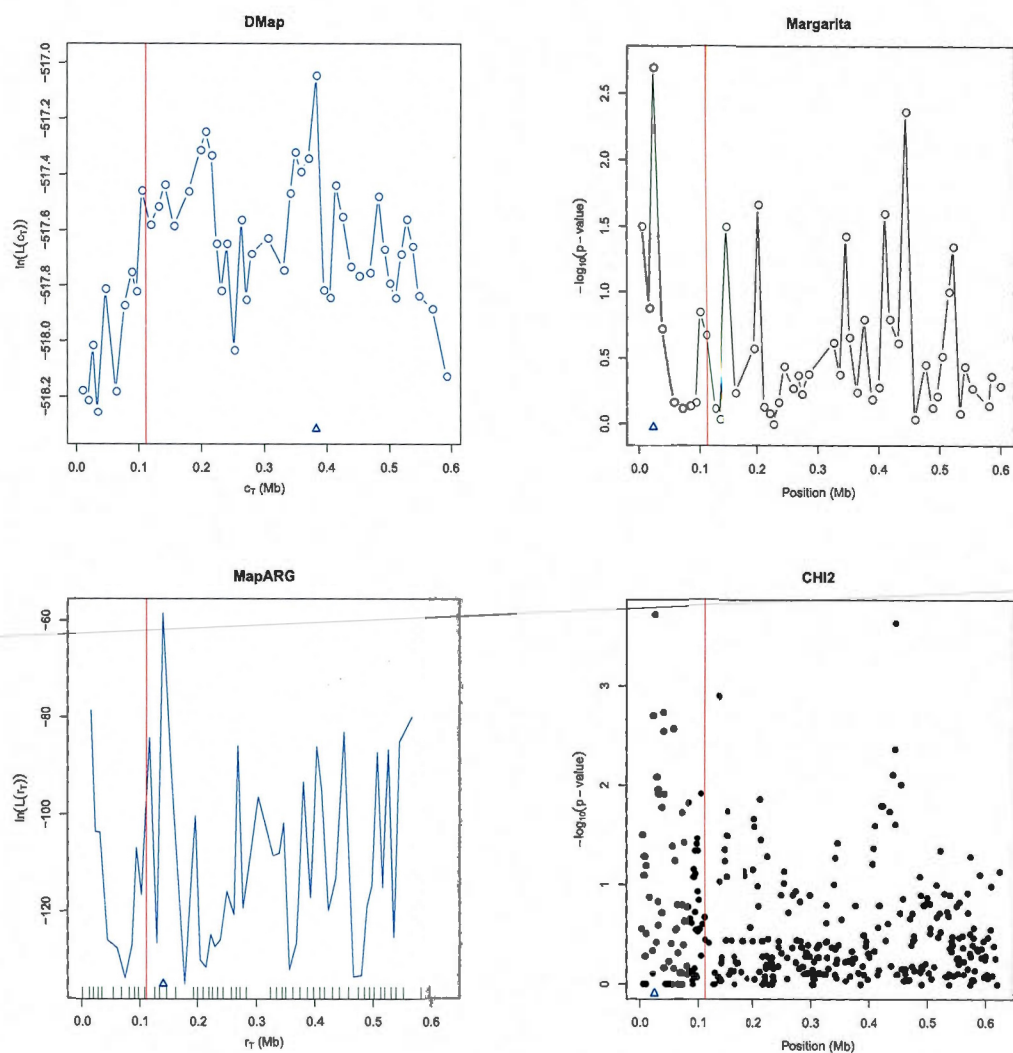


Figure 5.11 Figure illustrant les résultats obtenus avec les 4 méthodes de cartographie à comparer. L'échantillon généré selon le scénario G de la population 105 a été utilisé afin d'obtenir ces résultats.

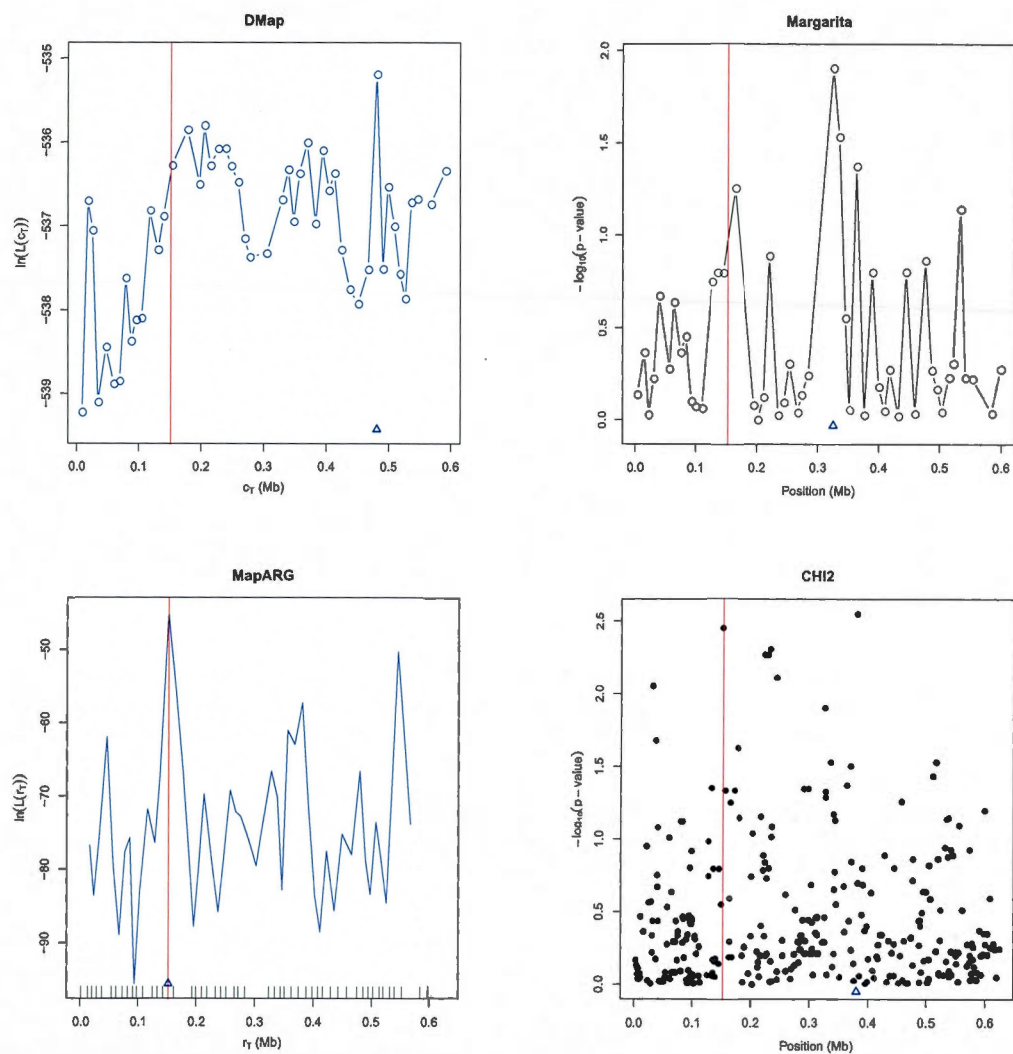


Figure 5.12 Figure illustrant les résultats obtenus avec les 4 méthodes de cartographie à comparer. L'échantillon généré selon le scénario H de la population 105 a été utilisé afin d'obtenir ces résultats.

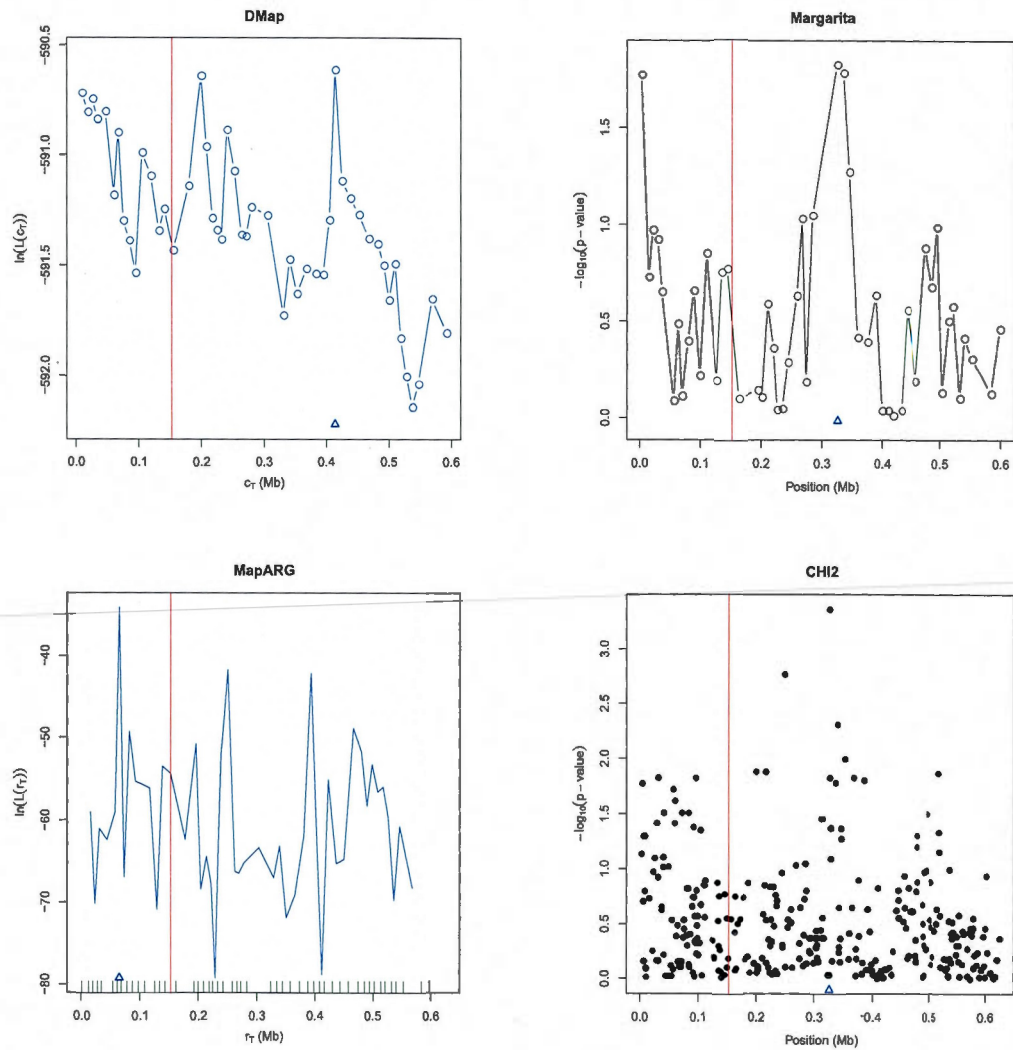


Figure 5.13 Figure illustrant les résultats obtenus avec les 4 méthodes de cartographie à comparer. L'échantillon généré selon le scénario I de la population 105 a été utilisé afin d'obtenir ces résultats.

5.3.3 Distribution des biais obtenus avec chaque méthode

Dans cette sous-section, nous avons comparé les résultats obtenus avec les méthodes DMap, MapARG, Margarita et Chi2. Pour ce faire, nous avons calculé le biais des estimateurs obtenus par chacune des méthodes lors de leur utilisation avec les échantillons provenant des scénarios A, B, C et D des populations 1 à 100 et avec les échantillons provenant des scénarios F, G, H et I des populations 101 à 200. Ceci nous a permis d'obtenir la distribution du biais pour chaque méthode et pour chaque scénario.

Les paramètres de simulation utilisés pour la méthode DMap ont été ceux décrit à la section 5.2.1. Pour la méthode Margarita, les paramètres de simulation utilisés sont exactement les mêmes que pour la méthode DMap, c'est-à-dire que le nombre de généalogies générés est de $M = 50$ et le nombre de marqueurs considérés par séquence est de $L = 64$ pour les scénarios A, B, C et D et de $L = 50$ pour les scénarios F, G, H et I. Quant à la méthode MapARG, nous avons utilisé le même nombre d'ARG simulés et de marqueurs par séquence que pour les méthodes DMap et Margarita, mais nous avons aussi utilisé la vraisemblance composite et conditionnelle introduite à la section 3.2.4. Le nombre de marqueurs par fenêtre qui a été utilisé est de $d - 1 = 2$. Finalement, nous avons utilisé tous les marqueurs contenus sur nos séquences génétiques pour la méthode Chi2, soit plus d'une centaine, car cette méthode étant très rapide, nous n'avons pas de contrainte au niveau du nombre de marqueurs que nous pouvons utiliser, et c'est ce qui se fait dans la pratique courante.

Les figures 5.14 à 5.21 présentent les résultats que nous avons obtenus pour chaque scénario (A à D et F à I) suite à ces simulations. Chacune de ces figures contient deux graphiques : celui de gauche représente la distribution du biais de chaque méthode à l'aide de diagrammes en boîte et celui de droite représente la distribution de la valeur absolue du biais de chaque méthode à l'aide de fonctions de répartition empirique, comme cela se fait habituellement dans la littérature en statistique génétique. La ligne pointillée présente sur les graphiques de droite illustre la fonction de répartition théorique que nous obtiendrions si l'estimation de la position du TIM était choisi aléatoirement.

Cette fonction de répartition est très utile, car elle nous permet de vérifier si les méthodes de cartographie génétique considérées sont meilleures qu'une méthode qui n'utilise pas les données ; nous pouvons donc vérifier si elles réussissent à utiliser correctement l'information contenue dans les données.

En observant les figures 5.14, 5.15 et 5.16, nous pouvons constater que les 4 méthodes semblent donner de bons résultats. En effet, pour chacun de ces 3 scénarios (A, B et C), la probabilité que le biais soit plus petit que 0.05 Mb est supérieure à 0.75, et ce, pour les 4 méthodes d'intérêt. Ce résultat n'est pas très surprenant, car dans les scénarios A, B et C, le risque relatif est assez élevé, nous nous attendions donc que les méthodes de cartographie génétique fonctionnent bien dans ces conditions. Il est cependant difficile de départager les méthodes, car nous n'observons pas de grandes différences entre elles ; nous pouvons toutefois affirmer que la méthode Margarita semble généralement donner de moins bons résultats que les autres, car la ligne rouge est la plupart du temps en dessous des autres lignes et que la méthode DMap semble mieux fonctionner que les autres méthodes pour les scénarios B et C.

La figure 5.17 présente les résultats obtenus avec le scénario D ; la maladie considérée dans ce scénario présente beaucoup de phénocopie (la probabilité f_0 est très élevée) et elle a un risque relatif très faible. Pour chacune des méthodes, nous pouvons observer que la variance des estimations est très grande et que la fonction de répartition empirique du biais est très près de la ligne pointillée. Les 4 méthodes ne semblent donc pas très bien fonctionner pour ce scénario.

À la figure 5.18, nous pouvons observer que les 4 méthodes donnent de bons résultats, bien que le risque relatif pour le scénario F soit relativement faible. La méthode DMap semble donner des résultats légèrement meilleurs que ceux des autres méthodes puisque la variance de ces estimations semblent être légèrement inférieure à celle des autres méthodes et que sa fonction de répartition empirique est quasiment toujours au dessus des fonctions des autres méthodes.

Finalement, les figures 5.19, 5.20 et 5.21 représentent les résultats obtenus avec les

scénarios G, H et I respectivement. Sur ces 3 figures, nous pouvons observer que les résultats obtenus par les 4 méthodes ne sont pas très bons. En effet, les résultats obtenus avec les scénarios G et H sont à peine meilleurs que ceux que nous obtiendrions avec un estimateur n'utilisant pas l'information contenue dans les données et les résultats obtenus avec le scénario I sont catastrophiques, car ils ne sont même pas meilleur que ceux que nous obtiendrions avec un estimateur n'utilisant pas l'information contenue dans les données. Ces résultats peuvent s'expliquer par le fait que les risques relatifs pour ces scénarios sont très faibles et peut-être aussi que la fréquence de maladie est faible pour les scénarios H et I. Notons toutefois que la méthode DMap fonctionne mieux que les autres méthodes pour le scénario G, ce qui est quand même encourageant.

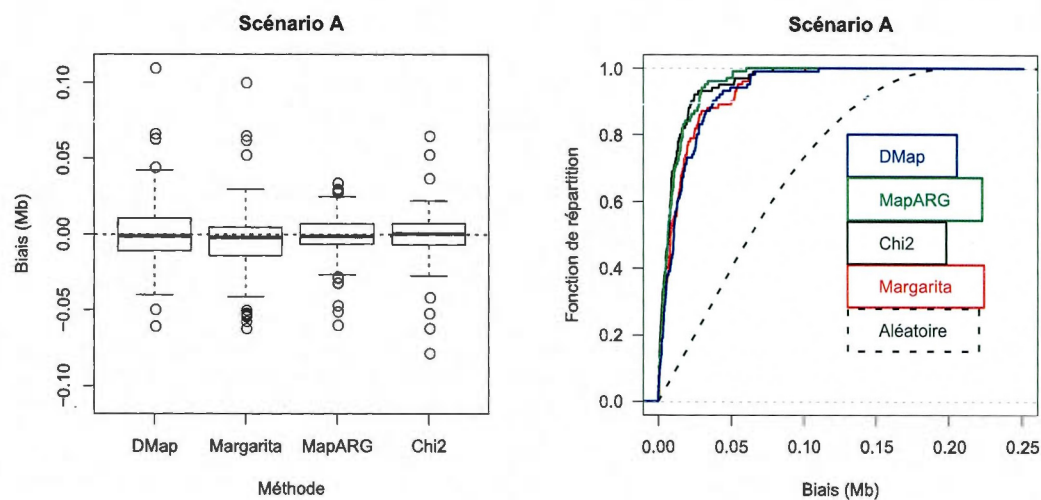


Figure 5.14 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario A.

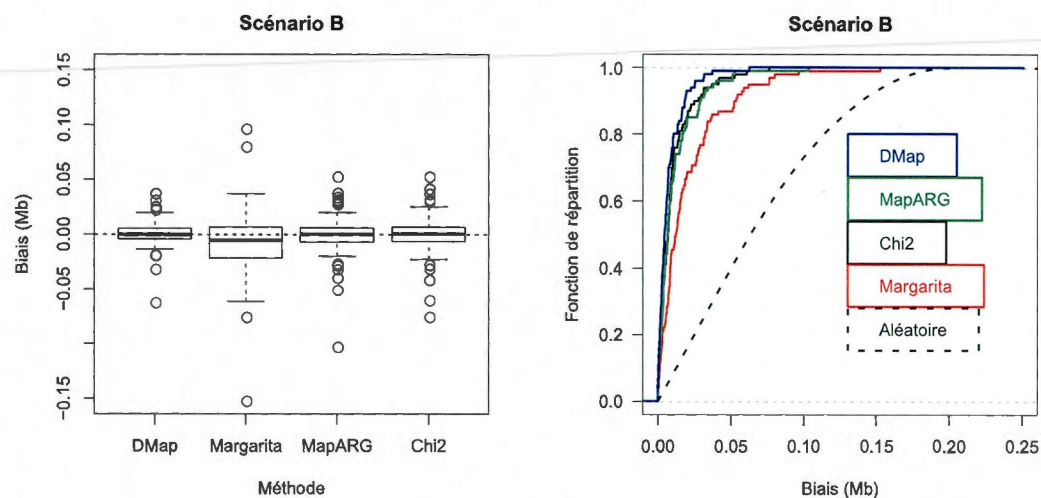


Figure 5.15 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario B.

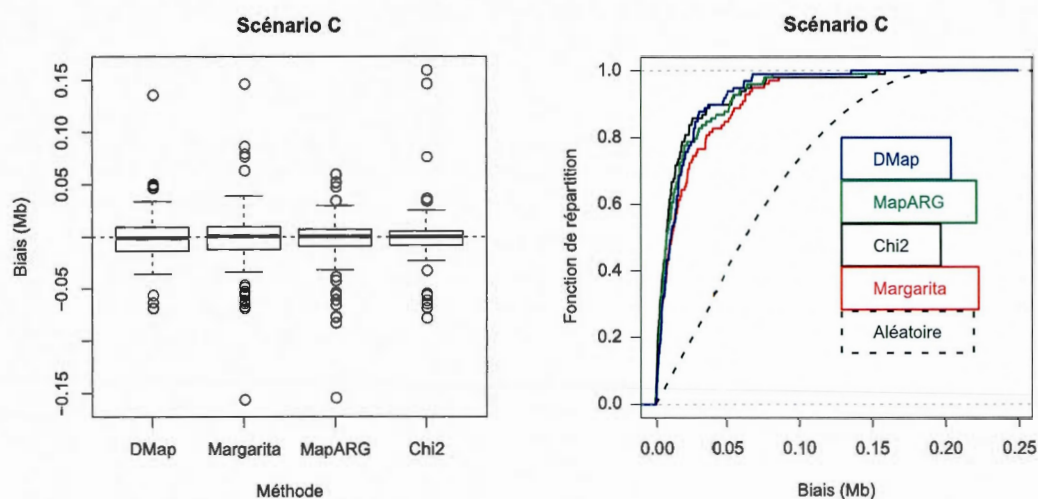


Figure 5.16 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario C.

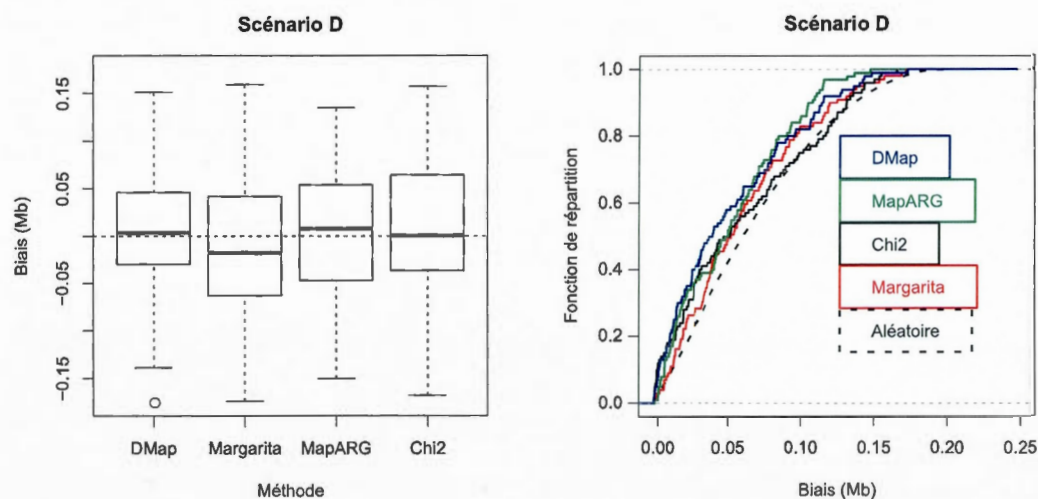


Figure 5.17 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario D.

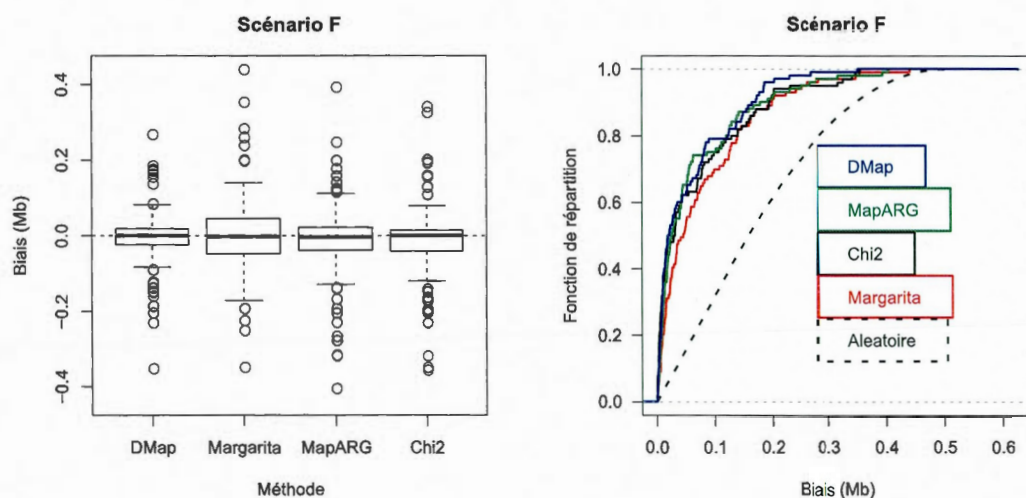


Figure 5.18 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario F.

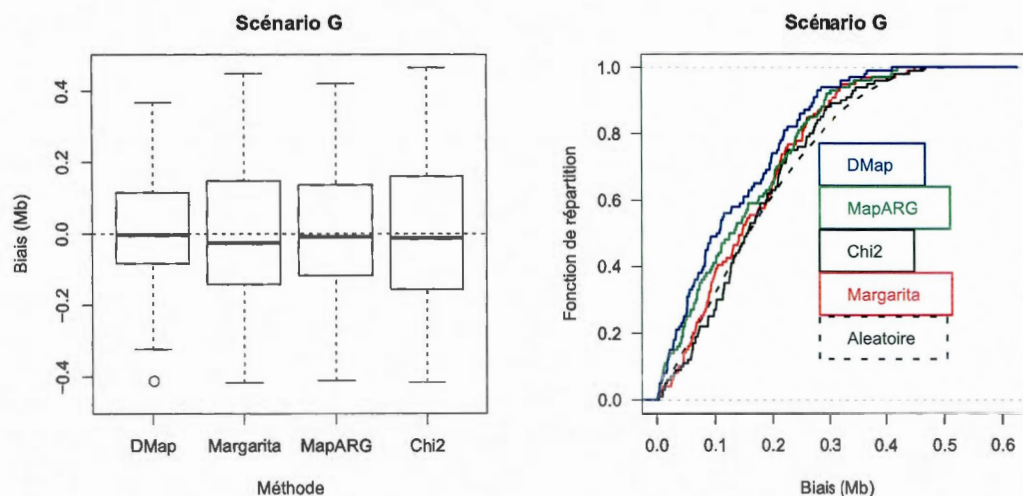


Figure 5.19 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario G.

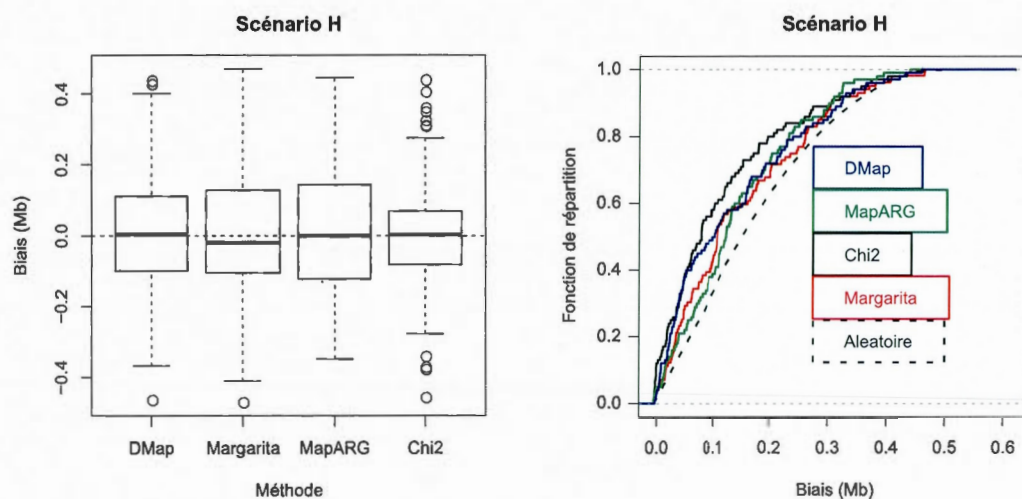


Figure 5.20 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario H.

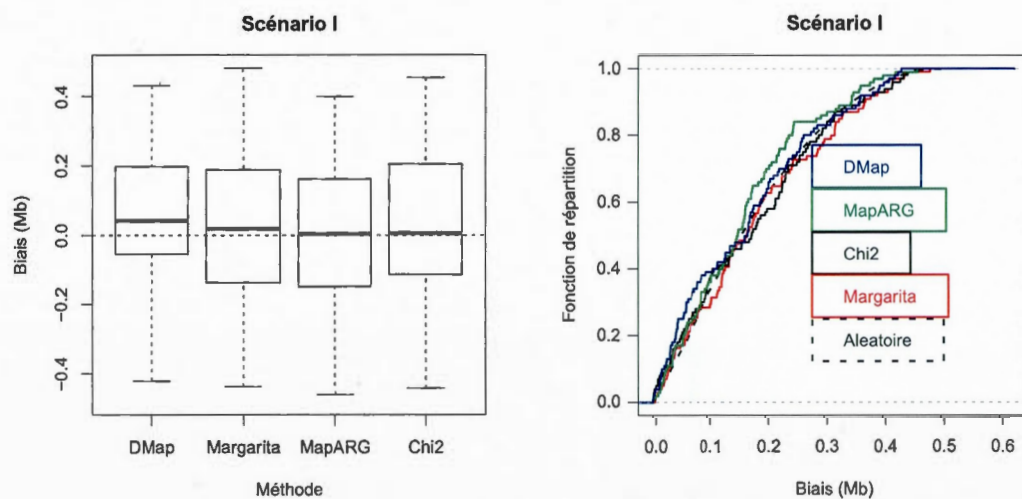


Figure 5.21 Figure illustrant la distribution du biais des estimateurs obtenus avec les 4 méthodes à comparer pour les échantillons générés selon le scénario I.

5.4 Discussion

Les résultats présentés dans la section précédente sont encourageants. En effet, DMap donne de bons résultats lorsque le risque relatif de la maladie considérée est élevé (scénario A, B et C), mais aussi lorsqu'il est plus faible (scénario F). Par contre, les résultats obtenus pour des modèles génétiques présentant de très faibles risques relatifs (scénario D, G, H et I) ne sont pas satisfaisants. En effet, dans ces situations, notre méthode, tout comme les méthodes MapARG, Margarita et Chi2, n'est pas bien meilleure qu'une méthode qui n'utiliserait aucune information provenant des données, il y a donc encore du travail à accomplir afin d'améliorer les performances de DMap. Il serait par exemple intéressant d'étudier l'effet du nombre de marqueurs (L) considérés lors de l'utilisation de DMap; nous apprendrions peut-être que les résultats seraient meilleurs si nous en utilisions un plus grand nombre. Nous devrions aussi, similairement à la méthode MapARG, travailler au développement de vraisemblance composite afin que la méthode DMap puisse utiliser des fenêtres de marqueurs. De plus, la distribution proposée $Q(\cdot)$ de Fearnhead et Donnelly semble construire de bonnes généalogies, cependant nous croyons que l'espace des graphes consistants avec un échantillon donné n'est peut-être pas bien exploré. Il serait donc intéressant d'étudier plus en détail cette distribution et de tenter de la modifier si nous découvrons qu'elle n'est pas adéquate.

La méthodologie que nous avons élaborée dans ce mémoire possède la qualité d'être flexible, c'est-à-dire qu'elle pourra facilement être adaptée dans le futur à des contextes de cartographie génétique fine plus général. En effet, il suffit d'apporter quelques modifications à la façon de calculer $P(\Phi|G, c_T)$ afin que DMap puisse être utilisée dans un contexte où le phénotype n'est plus qualitatif, mais bien quantitatif ou dans un contexte où l'on ne suppose plus que la fonction de pénétrance de la maladie d'intérêt est connue. Nous aimerions de plus pouvoir utiliser dans le futur la méthode DMap afin de travailler avec des maladies influencées par plus d'une mutation; nous croyons qu'il est possible de modifier l'algorithme de DMap sans trop de mal afin d'estimer l'emplacement de plus d'une mutation à la fois.

CONCLUSION

L'objectif de ce mémoire était de développer une nouvelle méthode de cartographie génétique fine, DMap, ayant la particularité de pouvoir être utilisée et d'être efficace en présence de modèles génétiques complexes. Nous voulions de plus tester la performance de notre nouvelle méthode et la comparer à la performance de quelques méthodes similaires de cartographie génétique .

Nous avons tout d'abord étudié trois méthodes de cartographies génétiques fine et nous avons ensuite utilisé ce qui nous semblait être les meilleurs éléments de chaque méthode afin de créer la méthode DMap. Similairement à la méthode MapARG, notre nouvelle méthode permet d'approximer la fonction de vraisemblance d'un paramètre de position (c_T) à l'aide de la théorie de la coalescence. Similairement à la méthode Margarita, l'inférence du génotype à un marqueur donné se fait en ajoutant une mutation sur l'arbre partiel du marqueur donné. Finalement, le calcul de la probabilité du vecteur contenant les phénotypes des séquences d'un échantillon se fait similairement à la méthode LATAG.

Nous avons de plus créé un programme en langage C++ permettant d'exécuter notre nouvelle méthode d'une façon rapide et efficace. Il nous est donc possible de faire des simulations en considérant plusieurs marqueurs par séquence sans avoir recours à l'usage de fenêtres de marqueurs, contrairement à la méthode MapARG.

Afin d'évaluer notre nouvelle méthode, nous l'avons comparée aux méthodes MapARG, Margarita et Chi2. Pour cela, nous avons simulé 800 échantillons provenant de 200 différentes populations de séquences génétiques. Pour chacun de ces échantillons, nous avons calculé le biais des estimations fournies par les différentes méthodes à comparer et nous avons ensuite illustré la distribution de ces biais à l'aide de diagrammes en boîtes et de fonctions de répartition estimées. Les résultats obtenus sont encourageants ;

DMap donne de bons résultats en présence de modèles génétiques simples et malgré le fait que les résultats ne sont pas tout à fait satisfaisants en présence de modèles génétiques complexes, ils sont du moins en général meilleurs que ceux obtenus avec d'autres méthodes similaires de cartographie génétique. Nous semblons donc être sur la bonne voie afin développer un outil performant de cartographie génétique fine.

APPENDICE A

APPROXIMATION DE LA DISTRIBUTION CONDITIONNELLE $\pi(\cdot|H)$

Les sections suivantes sont dédiées à la présentation détaillée de l'approximation de la distribution conditionnelle $\pi(\cdot|H)$ proposée par Fearnhead et Donnelly (2001) et introduite à la section 3.2.3 de ce mémoire.

A.1 Interprétation généalogique

Supposons que l'on connaisse k séquences présentes à une génération τ donnée d'un graphe de recombinaison ancestral et que l'on veuille construire la $(k+1)^{\text{e}}$ séquence présente dans cette même génération. Expliquons d'un point de vue généalogique ce que nous devons faire. On sait que les séquences présentes à un état d'une généalogie ont assurément un ancêtre commun, on est donc certain que les différents types de séquences rencontrés au cours d'une généalogie ont été créés par des recombinaisons et des mutations des séquences des générations précédentes. Ainsi, la $(k+1)^{\text{e}}$ séquence sera composée de fragments des k séquences de la génération τ , et ce afin de représenter les événements de recombinaison du passé. De plus, la transmission d'un fragment d'une séquence quelconque à la $(k+1)^{\text{e}}$ séquence peut se faire avec des erreurs, et ce afin de symboliser les événements de mutation du passé. On peut donc s'imaginer la $(k+1)^{\text{e}}$ séquence comme étant une mosaïque imparfaite des k séquences présentes à un état de notre graphe.

On obtient finalement que l'approximation $\hat{\pi}(\alpha|H)$ est la somme sur toutes les façons de

construire une séquence de type α à partir des séquences contenues dans l'échantillon H . Ceci signifie donc qu'il faut calculer toutes les façons possibles de découper la séquence α en fragments et d'assigner à chacun de ces fragments une séquence parmi celles déjà présentes dans l'échantillon.

A.2 Calcul formel de $\hat{\pi}(\cdot|H)$

Posons tout d'abord h_1, \dots, h_k pour les types des k séquences présentes dans l'échantillon H . Nous allons expliquer les calculs servant à l'obtention de la valeur $\hat{\pi}(h_{k+1}|h_1, \dots, h_k)$. Il faut tout d'abord mentionner que seuls les marqueurs ancestraux de la séquence h_{k+1} seront considérés lors de ce calcul. En effet, le matériel non-ancestral n'apporte aucune information afin de savoir si la séquence est près ou éloignée, d'un point de vue génétique, des autres séquences de l'échantillon, on peut donc l'ignorer. Notons par s le nombre de marqueurs pour lesquels la séquence h_{k+1} possède du matériel ancestral. Le calcul de $\hat{\pi}(\cdot|H)$ est basé sur le fait qu'il n'y a pas de matériel non-ancestral parmi les séquences h_1, \dots, h_k ; il faut donc assigner un allèle «0» ou «1» à tous les marqueurs non-ancestraux présents dans l'échantillon. Ceci est fait de la façon suivante : on assigne un allèle a à un marqueur m avec une probabilité, notée $p_{a,m}$, égale à la proportion des séquences ayant un allèle a au marqueur m parmi les séquences possédant du matériel ancestral à ce marqueur.

Nous avons vu à la section précédente que l'allèle sur chaque marqueur m de la séquence h_{k+1} est une copie, possiblement erronée, du m^{e} marqueur d'une des k séquences de l'échantillon. Posons X_m comme étant le numéro de la séquence dont on a copié l'allèle m sur h_{k+1} . Par exemple, pour une séquence h_{k+1} possédant quatre marqueurs ancestraux, on pourrait obtenir $(X_1, X_2, X_3, X_4) = (2, 2, 4, 5)$, ce qui signifierait qu'on a découpé notre séquence en 3 fragments et qu'il y a donc eu 2 recombinaisons; une entre les marqueurs 2 et 3 et une autre entre les marqueurs 3 et 4. Il est possible de voir la suite (X_1, \dots, X_s) comme étant une chaîne de Markov dont les états sont les k séquences de l'échantillon avec $P(X_1 = x) = 1/k, \forall x \in \{1, 2, \dots, k\}$ et ayant comme probabilités de

transition :

$$P(X_{m+1} = x | X_m = y) = \begin{cases} \exp\left(\frac{\rho_m}{k+\rho_m}\right) + \left(1 - \exp\left(\frac{\rho_m}{k+\rho_m}\right)\right) \frac{1}{k} & \text{si } x = y, \\ \left(1 - \exp\left(\frac{\rho_m}{k+\rho_m}\right)\right) \frac{1}{k} & \text{sinon,} \end{cases}$$

où $\rho_m = 4Nr_m$ est le taux de recombinaison entre les marqueurs m et $m+1$ et $\rho_m/(k+\rho_m)$ est le taux de recombinaison entre les marqueurs m et $m+1$ pour un échantillon de k séquences ; ce taux découle directement de la formule d'échantillonnage d'Ewens (Ewens, 1972). Notons qu'à chaque fois qu'une recombinaison survient, c'est-à-dire qu'on change d'état dans notre chaîne de Markov, le nouvel état est choisi aléatoirement parmi les k séquences de l'échantillon incluant la séquence que l'on vient de quitter.

Introduisons maintenant les probabilités de copier un marqueur d'une façon exacte ou erronée. Grâce à la formule d'échantillonnage d'Ewens, nous savons que la probabilité de copier une séquence avec une erreur, c'est-à-dire qu'il y ait une mutation, est de $\theta/(k+\theta)$. Dans le contexte d'un modèle des sites infinis pour l'apparition des mutations, on obtient que si une mutation est possible au m^e marqueur, alors

$$P(h_{k+1,m} = a | X_m = x, h_1, \dots, h_k) = \begin{cases} k/(\theta + k) & \text{si } h_{x,m} = a, \\ \theta/(k + \theta) & \text{sinon,} \end{cases}$$

où $h_{x,m}$ représente le m^e marqueur sur la séquence x . S'il n'y a pas de mutation possible au m^e marqueur, alors

$$P(h_{k+1,m} = a | X_m = x, h_1, \dots, h_k) = \begin{cases} 1 & \text{si } h_{x,m} = a, \\ p_{a,m} \frac{\theta}{(k+\theta)} + p_{1-a,m} \frac{k}{(\theta+k)} & \text{si } h_{x,m} \text{ est non ancestral.} \\ 0 & \text{sinon,} \end{cases}$$

où $p_{a,m}$ est la probabilité d'assigner l'allèle a au marqueur m . On calcule finalement $\hat{\pi}(h_{k+1}|h_1, \dots, h_k)$ en sommant sur toutes les possibilités de X_m pour $m = 1, \dots, s$. Ce calcul se fait aisément de façon récursive.

A.3 Calcul récursif de $\hat{\pi}(h_{k+1}|h_1, \dots, h_k)$

Posons $\alpha_m(x) = P(h_{k+1, \leq m}, X_m = x)$, où $h_{k+1, \leq m}$ représente le type des m premiers marqueurs de l'haplotype h_{k+1} . On peut calculer $\alpha_1(x)$ directement pour $x \in$

(h_1, \dots, h_k) de la façon suivante :

$$\alpha_1(x) = 1/k \cdot \text{Pr}(h_{k+1,1} = a | X_1 = x, h_1, \dots, h_k),$$

où a est l'allèle au premier marqueur à évaluer de la séquence h_{k+1} . On peut ensuite calculer $\alpha_2(x), \dots, \alpha_s(x)$ pour $x \in (h_1, \dots, h_k)$ d'une façon récursive :

$$\begin{aligned} \alpha_{m+1}(x) &= \gamma_{m+1}(x) \cdot P(X_{m+1} = x | X_m = x') \cdot \sum_{x'=1}^k \alpha_m(x') \\ &= \gamma_{m+1}(x) \cdot \left((1 - p_m) \alpha_m(x) + p_m \frac{1}{k} \sum_{x'=1}^k \alpha_m(x') \right), \end{aligned}$$

où $p_m = \exp(\rho_m / (k + \rho_m))$ et $\gamma_{m+1}(x) = P(h_{k+1,m+1} | X_{m+1} = x, h_1, \dots, h_k)$.

On obtient finalement que $\hat{\pi}(h_{k+1} | h_1, \dots, h_k) = \sum_{x=1}^k \alpha_s(x)$.

APPENDICE B

ALGORITHME EN LANGAGE C++

Cette annexe contient le code écrit en langage C++ permettant d'extraire un arbre partiel A_{π_z} d'un graphe G à partir de l'algorithme présenté à la section 4.3.2. Les ensembles d'intérêt $N_{A_{\pi_z}}$, $T_{A_{\pi_z}}$ et $D_{A_{\pi_z}}$ y sont obtenus d'une façon itérative.

```
double ExtraireArbrePartiel (double piZ, VVecD grapheG, VecD Temps) {  
  
    /* La variable piZ contient la position du marqueur dont on veut extraire l'arbre  
    partiel, la matrice grapheG contient tous les événements survenus dans le graphe G  
    en ordre chronologique et le vecteur Temps contient les temps d'attente entre chaque  
    événements de la matrice grapheG. */  
  
    VecI N_A_piZ; // Vecteur contenant les numéros des noeuds inclus dans l'arbre partiel  
    A $_{\pi_z}$ .  
  
    VecD tempsTempo; // Vecteur contenant les temps auxquels chaque séquence a été im-  
    pliquée dans un événement pour la dernière fois.  
  
    VecD T_A_piZ; // Vecteur contenant la longueur de la branche qui lie chaque noeud  
    du vecteur N_A_piZ à son parent.  
  
    VVecI D_A_piZ; // Matrice dont chaque ligne représente un noeud contenu dans N_A_piZ.  
    Pour chacun de ses noeuds, la première colonne contient le nombre de ses descendants  
    parmi les feuilles du graphe qui sont des cas et la deuxième colonne contient le nombre  
    de ses descendants parmi les feuilles du graphe qui sont des témoins.
```

double TempsCumulatif = 0.0; // Variable itérative contenant le temps entre la première génération de notre graphe et une génération i du graphe G .

int num1, num2, num3, num4; // Variables contenant des valeurs temporaires facilitant la compréhension du code.

// On initialise les vecteurs N_A_piZ, T_A_piZ, tempsTempo et la matrice D_A_piZ pour les séquences contenus dans notre échantillon.

```
for(int i = 0; i != tailleEchantillon; i++){
    N_A_piZ.push_back(i);
    T_A_piZ.push_back(0);
    tempsTempo.push_back(0);
    if(Pheno[i]=="cas"){
        D_A_piZ.push_back(1,0);
    } else {
        D_A_piZ.push_back(0,1);
    }
}
```

// Pour chaque événement du graphe G , on met à jour les vecteurs N_A_piZ, T_A_piZ et la matrice D_A_piZ.

```
for(int i = 0; i != grapheG.size(); i++){
    TempsCumulatif = TempsCumulatif + Temps[i];

    if(grapheG[i][0]==1 || grapheG[i][0]==2) { // L'événement est une coalescence
        num1 = 0; num2 = 0; num3 = 0; num4 = 0;

        // On vérifie si le 1er noeud qui coalesce fait partie du vecteur N_A_piZ
        int ligne1 = ChercherLigneImplique(grapheG[i][1], N_A_piZ);
```

```

// Si oui, ses descendants font partie des descendants du noeud résultant:
if(ligne1 != -1){
    num1 = D_A_piZ[ligne1][0];
    num2 = D_A_piZ[ligne1][1];
    T_A_piZ[ligne1] = TempsCumulatif - tempsTempo[ligne1];
}

// On vérifie si le 2eme noeud qui coalesce fait partie du vecteur N_A_piZ
int ligne2 = ChercherLigneImplique(grapheG[i][2], N_A_piZ);
// Si oui, ses descendants font partie des descendants du noeud résultant:
if(ligne2 != -1){
    num3 = D_A_piZ[ligne2][0];
    num4 = D_A_piZ[ligne2][1];
    T_A_piZ[ligne2] = TempsCumulatif - tempsTempo[ligne2];
}

// Si au moins un des deux noeuds qui coalescent fait partie de l'arbre
// partiel à extraire alors le noeud résultant en fait aussi partie. On
// doit donc mettre à jour les vecteurs et la matrice d'intérêt.
if(ligne1 != -1 || ligne2 != -1){
    //On ajoute le nouveau noeud de l'arbre.
    N_A_piZ.push_back(grapheG[i][3]);

    // On ajoute une ligne à la matrice D_A_piZ pour le nouveau noeud.
    D_A_piZ.push_back(num1+num3,num2+num4);

    // On met à jour les vecteurs des longueurs de branches.
    tempsTempo.push_back(TempsCumulatif);
    T_A_piZ.push_back(0);
}

```

```

}

if(grapheG[i][0]==3) { // L'événement est une mutation

    // On vérifie si le noeud qui mute fait partie du vecteur N_A_piZ
    int ligne1 = ChercherLigneImplique(grapheG[i][1], N_A_piZ);
    // Si oui, alors le noeud résultant en fait aussi partie et ses descendants
    // sont exactement les mêmes que ceux du noeud qui mute:
    if(ligne1 != -1){
        //On ajoute le nouveau noeud de l'arbre.
        N_A_piZ.push_back(grapheG[i][2]);

        // On ajoute une ligne à la matrice D_A_piZ pour le nouveau noeud.
        D_A_piZ.push_back(D_A_piZ[ligne1][0],D_A_piZ[ligne1][1]);

        // On met à jour les vecteurs des longueurs de branches.
        tempsTempo.push_back(TempsCumulatif);
        T_A_piZ[ligne1] = TempsCumulatif - tempsTempo[ligne1];
        T_A_piZ.push_back(0);
    }
}

if(grapheG[i][0]==4) { // L'événement est une recombinaison

    // On vérifie si le noeud qui recombine fait partie du vecteur N_A_piZ
    int ligne1 = ChercherLigneImplique(grapheG[i][1], N_A_piZ);
    // Si oui, alors un des deux noeuds résultants en fait aussi partie et
    // ses descendants sont exactement les mêmes que ceux du noeud qui
    // recombine:
    if(ligne1 != -1){

```

```

// On doit ajouter à notre arbre partiel le parent ayant un marqueur
// ancestral à la position piZ suite à cette recombinaison.

if(piZ < grapheG[i][4]){
    N_A_piZ.push_back(grapheG[i][2]);
} else {
    N_A_piZ.push_back(grapheG[i][3]);
}

// Les descendants du nouveau noeud sont exactement les mêmes
// que ceux du noeud qui recombine.
D_A_piZ.push_back(D_A_piZ[ligne1][0], D_A_piZ[ligne1][1]);

// On met à jour les vecteurs des longueurs de branches.
tempsTempo.push_back(TempsCumulatif);
T_A_piZ[ligne1] = TempsCumulatif - tempsTempo[ligne1];
T_A_piZ.push_back(0);
}
}
}

```


RÉFÉRENCES

- Boucher, G. 2009. « Intégration de la réalité diploïde et des modèles de pénétrance à une méthode de cartographie génétique fine ». Mémoire de maîtrise, Université du Québec à Montréal.
- Campbell, N. A., et R. Mathieu. 1995. *Biologie*. Éditions du Renouveau Pédagogique.
- Churchill, G., et R. Doerge. 1994. « Empirical threshold values for quantitative trait mapping ». *Genetics*, vol. 138, p. 963–971.
- Elston, R. C. 2000. « Introduction and overview, statistical methods in genetic epidemiology ». *Statistical methods in medical research*, vol. 9, no. 6, p. 527–541.
- Fearnhead, P., et P. Donnelly. 2001. « Estimating recombination rates from population genetic data ». *Genetics*, vol. 159, no. 3, p. 1299–1318.
- Forest, M. 2010. « Cartographie génétique fine simultanée de deux gènes ». Mémoire de maîtrise, Université du Québec à Montréal.
- Griffiths, R. C., et P. Marjoram. 1996. « Ancestral inference from samples of dna sequences with recombination ». *Journal of Computational Biology*, vol. 3, no. 4, p. 479–502.
- Griffiths, R. C., et S. Tavaré. 1994a. « Ancestral inference in population genetics ». *Statistical Science*, vol. 9, no. 3, p. 307–319.
- . 1994b. « Simulating probability distributions in the coalescent ». *Theoretical population biology*, vol. 46, p. 131–159.
- Hein, J., M. H. Schierup et C. Wiuf. 2005. *Gene Genealogies, Variation and Evolution : A Primer in Coalescent Theory*. Oxford University Press.
- Hudson, R. R. 1991. « Gene genealogies and the coalescent process ». *Oxford Surveys in Evolutionary Biology*, vol. 7, p. 1–44.
- . 2001. « Two-locus sampling distributions and their application ». *Statistics*, vol. 159, p. 1805–1817.
- . 2002. « Generating samples under a wright-fisher neutral model of genetic variation ». *Bioinformatics*, vol. 18, p. 337–338.
- Jenkins, P. A. 2008. « Importance sampling on the coalescent with recombination ». Thèse de Doctorat, University of Oxford.

- Kingman, J. F. C. 1982. « The coalescent ». *Stochastic Processes and their Application*, vol. 13, p. 235–248.
- Lander, E. S., et N. J. Schork. 1994. « Genetic dissection of complex traits ». *Science*, vol. 265, no. 5181, p. 2037–2048.
- Larribe, F. 2003. « Cartographie génétique fine par le graphe de recombinaison ancestral ». Thèse de Doctorat, Université de Montréal.
- Larribe, F., et P. Fearnhead. 2011. « On composite likelihoods in statistical genetics ». *Statistica Sinica*, vol. 21, no. 1.
- Larribe, F., et S. Lessard. 2008. « A composite-conditional-likelihood approach for gene mapping based on linkage disequilibrium in windows of marker loci ». *Statistical applications in genetics and molecular biology*, vol. 7, p. Article 27.
- Larribe, F., S. Lessard et N. J. Schork. 2002. « Gene mapping via the ancestral recombination graph ». *Theoretical population biology*, vol. 62, no. 2, p. 215–229.
- Li, N., et M. Stephens. 2003. « Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data ». *Genetics*, vol. 165, p. 2213–2233.
- Minichiello, M. J., et R. Durbin. 2006. « Mapping trait loci by use of inferred ancestral recombination graphs ». *The American Journal of Human Genetics*, vol. 79, p. 910–922.
- Nordborg, M. 2007. *Handbook of Statistical Genetics*, chapitre 2, p. 843–877. John Wiley and Sons, Ltd, 3 édition.
- Olson, J. M., J. S. Witte et R. C. Elston. 1999. « Genetic mapping of complex traits ». *Statistics in medicine*, vol. 18, no. 21, p. 2961–2981.
- Stephens, M., et P. Donnelly. 2000. « Inference in molecular population genetics ». *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 62, no. 4, p. 605–655.
- Varin, C. 2008. « On composite marginal likelihoods ». *Advances in Statistical Analysis*, no. 92, p. 1–28.
- Zöllner, S., et J. K. Pritchard. 2005. « Coalescent-based association mapping and fine mapping of complex trait loci ». *Genetics*, vol. 169, no. 2, p. 1071–1092.